

Supporting Researchers:  
Analyzing  
the Scholarly Publication Life Cycle  
and Social Bookmarking Systems

Dissertation for the acquisition of the academic degree  
Doktor der Naturwissenschaften (Dr. rer. nat.)

Submitted to the Faculty of Electrical Engineering  
and Computer Science of the University of Kassel

By Stephan Doerfel

Submitted: Kassel, April 27, 2016

Defended: Kassel, September 08, 2016



## Erklärung

Hiermit versichere ich, dass ich die vorliegende Dissertation selbstständig, ohne unerlaubte Hilfe Dritter angefertigt und andere als die in der Dissertation angegebenen Hilfsmittel nicht benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen sind, habe ich als solche kenntlich gemacht. Dritte waren an der inhaltlich-materiellen Erstellung der Dissertation nicht beteiligt; insbesondere habe ich hierfür nicht die Hilfe eines Promotionsberaters in Anspruch genommen. Kein Teil dieser Arbeit ist in einem anderen Promotions- oder Habilitationsverfahren verwendet worden.

Stephan Doerfel



## Danksagung

Zuallererst gehört mein Dank Prof. Dr. Gerd Stumme für seine Unterstützung und die Betreuung meiner Dissertation, für viele gemeinsame Publikationen, für hilfreiche und wegweisende Gespräche, für ein sicheres und angenehmes Arbeitsklima an unserem Fachgebiet und für das mir entgegengebrachte Vertrauen.

Sehr herzlich danke ich Prof. Dr. Isabella Peters für die Übernahme der Begutachtung dieser Dissertation sowie für ihr konstruktives und hilfreiches Feedback, das sehr geholfen hat, diese Arbeit weiter zu verbessern.

Unvorstellbar wäre der Arbeitsalltag ohne meine Kollegen und Freunde am Fachgebiet Wissensverarbeitung. Für viele wertvolle Diskussionen, Tipps, technische Unterstützung, Motivationshilfe und entspannende Pausen mit Kickern oder Kuchen danke ich Andreas Schmidt, Beate Navarro Bullock, Björn Fries, Björn-Elmar Macek, Christoph Scholz, Daniel Zoller, Dominik Benz, Jens Illig, Jürgen Müller, Mark Kibanov, Martin Atzmüller, Monika Vopicka, Sebastian Böttger, Sven Stefani und Tom Hanika. Ganz besonders danke ich meinen Büronachbarn Andreas Hotho und Folke Mitzlaff, sowie meinem Post-Doc Robert Jäschke.

Jedes Kapitel dieser Arbeit enthält Beiträge aus gemeinsamen Veröffentlichungen und ich danke meinen Koautoren für diese erfolgreichen Kooperationen. Mein Dank gilt auch dem gesamten BibSonomy-Team. Nur durch die jahrelange, unermüdliche Arbeit zur Entwicklung und Verbesserung unseres Systems konnten die Daten entstehen, die mehreren Kapiteln dieser Dissertation zu Grunde liegen.

Unterstützt wurde ich auch durch die Deutsche Forschungsgemeinschaft (DFG), durch die Finanzierung der Projekte “Vernetzte Repositorien: Akademisches Publikationsmanagement / PUMA” sowie “Informationelle Selbstbestimmung im Web 2.0”, in denen einige Studien dieser Dissertation entstanden sind.

Neben der wissenschaftlichen Herausforderung stellt die Dissertation auch zahlreiche persönliche Herausforderungen. Mein ganz besonderer Dank gilt meiner Frau Ann-Kristin, die mich zu jeder Zeit uneingeschränkt unterstützt und motiviert hat und die mir stets zur Seite stand.



## Abstract

Researchers must face the exponential growth of the body of available scholarly literature, which makes it ever harder to keep track with one's own community, especially for newcomers. In this thesis, we explore different means of supporting researchers with that task. For this purpose, we follow two approaches: We provide analyses of research communities and of researchers' interactions through data that can be obtained from the phases in the life cycle of scholarly publications (creation, dissemination, usage, and citation in other publications). The resulting statistics and visualizations allow researchers to better understand their own communities, to identify the most important players and publications, and to find valuable conversational partners at conferences. For the analysis of publication usage and connections to citations, we turn to social bookmarking systems and investigate the actions of users in BibSonomy. The provided insights can help operators of such systems improve them. Our second approach is more proactive, focusing on supporting researchers by pointing them directly to important publications – through automatically computed personalized recommendations and through social peer review.

The analysis of research and researchers often relied on studying scholarly publications and their metadata. Such studies can reveal insights into how scientific work is conducted, they can shed light on communities and research topics, and they allow the measurement of certain forms of impact, a publication, an individual researcher, or a venue had. The exploited data – publication metadata– is generated when publications are created. The life cycle of a scholarly publication, however, just begins with a publication's *creation*: Publications are *disseminated* (e.g., presented at conferences), they are *used* (e.g., acquired, stored, collected, marked as to-read, and, of course, read), and they are *cited*. With the advent of the Web 2.0, traces of the activities in these phases have become observable. In this thesis, we collect and analyze datasets from all four stages of the publication life cycle. We thus go beyond traditional means of scientometrics, touching such fields as altmetrics, web log analysis, and role discovery. We not only present new insights into communities that have not been investigated before, but we also demonstrate new means of analysis that are generalizable to other communities as well. Among them are formal concept analysis to visualize influences between groups of authors and social network analyses of interaction networks. Our datasets comprise – next to a traditional publication corpus containing metadata and references – a face-to-face contact network, gathered from real-live interactions of researchers during a conference, and datasets from the scholarly social bookmarking system BibSonomy.

Social bookmarking services allow their users to publicly store and annotate resources, like web links, photos, videos, or publications. As representatives of the Web 2.0, social

---

bookmarking systems have attracted the interest of the research community. Through the central feature, tagging of resources, users of such systems create a data structure called folksonomy, in which users, resources, and tags are connected. The resulting network allows users to navigate between these folksonomic entities. In scholarly bookmarking systems, users store and manage publications. Thus, such systems are an ideal candidate for the investigation of publication usage. In this thesis, we study data of the popular system BibSonomy to address various aspects of the use of social bookmarking systems and the therein stored resources. Moreover, we analyze the usefulness for altmetrics by studying correlations between the usage of a publication and its citations, as well as predictive power of usage-features over future citations.

Scholarly bookmarking tools support researchers in their daily work with publications and their metadata. Still, the sheer number of available publications and its ever faster growth make it difficult to keep track of the relevant developments in one's field of research – an instance of the information overload problem. Therefore, recommendation systems can be employed to point users to particular publications using personalized ranking algorithms. Usually, such algorithms exploit information in user profiles, for instance, previously stored resources and the according tags, as well as information about similarity between entities or about their positions within the network of entities (the folksonomy) to recommend new items that the active user might find interesting. Similarly, a recommender can also assist the process of tagging by recommending suitable tags to users while they create a new post for some resource. We use the scenario of tag recommendation to thoroughly analyze the typical evaluation setup of folksonomic recommender systems using so-called graph-cores. We improve the setup by introducing a new, more flexible type of core to circumvent a structural drawback of the graph-core approach. We also point to several pitfalls of using cores for benchmarking recommendation algorithms. Moreover, we employ the scenario of resource recommendation – specifically the recommendation of scholarly publications – to investigate different ways of integrating publication metadata into the popular and versatile folksonomic recommendation algorithm FolkRank.

Finally, any tool that is offered on the web must comply with the law and its use must be socially compatible. Particularly difficult is the case of publicly visible ratings, where products are judged by users. For instance, in the case where resources are scholarly publications and thus the products of researchers (the authors), improper criticism may have consequences for researchers' careers or for decisions about funding allocation. Based on requirements that have been derived from German law, we describe and discuss opportunities and risks of social web systems in which users share, debate, and rate scholarly publications.

Altogether, this thesis relies on data from the scholarly publication life cycle to gain insights into research communities and the interaction of researchers with literature. We focus on social bookmarking systems, which reveal traces of its users' behavior and which provide a suitable tool to support researchers in their work with literature. Our contributions aim at supporting researchers in their work, as members of their respective communities and as producers and consumers of scholarly literature.



## Zusammenfassung

Durch das exponentielle Wachstum der Menge veröffentlichter wissenschaftlicher Literatur wird es für ForscherInnen immer schwieriger, einen Überblick über ihre Wissenschaftscommunity zu bekommen und zu behalten. In dieser Arbeit untersuchen wir verschiedene Mittel um ForscherInnen bei dieser Aufgabe zu unterstützen. Dabei folgen wir zwei Ansätzen: Zuerst analysieren wir Wissenschaftscommunities, Interaktionen zwischen WissenschaftlerInnen und die Nutzung von Literatur. Die Untersuchungen basieren auf Daten aus dem Lebenszyklus wissenschaftlicher Publikationen, der die Phasen der Erstellung, Verbreitung, Nutzung und Zitation einer Publikation umfasst. Die Statistiken und Visualisierungen, die aus solchen Untersuchungen entstehen, unterstützen ForscherInnen dabei, ihre eigenen Communities besser zu verstehen, die wichtigsten Akteure und Publikationen zu identifizieren oder auf Konferenzen interessante GesprächspartnerInnen zu finden. Für die Analyse der Nutzung und Zitierung von Publikationen betrachten wir soziale Verschlagwortungssysteme (Social Bookmarking Systems) und untersuchen dort die Aktionen der BenutzerInnen von BibSonomy. Die gewonnenen Erkenntnisse ermöglichen es BetreiberInnen solcher Systeme, diese entsprechend zu verbessern oder zu erweitern. Mit unserem zweiten, pro-aktiveren Ansatz wenden wir uns der direkteren Unterstützung von ForscherInnen beim Umgang mit Literatur zu: Mittels automatischer Empfehlungsverfahren können WissenschaftlerInnen direkt auf für sie relevante Publikationen aufmerksam gemacht werden. Im sogenannten Social Peer Review veröffentlichen LeserInnen Kritiken um anderen ihre Einschätzung zur Qualität einer Publikation zugänglich zu machen.

Die Analyse von Wissenschaftscommunities und Wissenschaftsfeldern stützt sich häufig auf wissenschaftliche Publikationen und deren Metadaten. Solche Studien gewähren Einblicke in Communities und sie ermöglichen die Messung von bestimmten Formen des Einflusses, bzw. der Relevanz von Veröffentlichungen, einzelnen ForscherInnen oder einer Publikationsplattform. Die dabei verwendeten Publikationsmetadaten entstehen, wenn wissenschaftliche Beiträge geschrieben und veröffentlicht werden. Das Veröffentlichen einer Publikation stellt jedoch nur den Beginn von deren Lebenszyklus dar. Publikationen werden verbreitet (unter anderem auf Konferenzen vorgestellt), können genutzt (beispielsweise gekauft, gespeichert, gesammelt, zum Lesen vorgemerkt und natürlich gelesen), und können zitiert werden. Mit dem Aufkommen des Web 2.0 sind Spuren dieser Aktivitäten sichtbar und messbar geworden. In dieser Dissertation sammeln und analysieren wir Datensätze aus allen vier Phasen des Publikationslebenszyklus (Erstellung, Verbreitung, Nutzung und Zitation). Wir gehen somit über die traditionellen Analysen der Szientometrie hinaus und befassen uns unter anderem mit Altmetrics, Web-Log-Analyse und der Entdeckung von Rollen in sozialen Netzwerken. Wir präsentieren neue Einsichten in Communities, die zuvor nicht

---

untersucht wurden, und verwenden dabei Analysemethoden, die leicht auch auf andere Communities übertragbar sind. Unter anderem nutzen wir Formale Begriffsanalyse um Einflüsse zwischen Gruppen von AutorInnen sichtbar zu machen, sowie Graphanalysen um Interaktionsnetzwerke von WissenschaftlerInnen auszuwerten. Unsere Datensätze umfassen – neben einem traditionellen Publikationskorpus mit Metadaten und Referenzen – ein Netzwerk von persönlichen Gesprächskontakten zwischen ForscherInnen im Rahmen einer Konferenz, sowie Datensätze aus dem wissenschaftlichen sozialen Verschlagwortungssystem BibSonomy.

In sozialen Verschlagwortungssystemen speichern und annotieren NutzerInnen Ressourcen öffentlich, z. B. Web-Links, Fotos, Videos oder Publikationen. Als Vertreter des Web 2.0 haben diese Systeme großes Interesse in der Forschungscommunity auf sich gezogen. Durch das zentrale Feature – das Annotieren (Taggen) von Ressourcen – erschaffen die NutzerInnen gemeinsam eine Datenstruktur, die sogenannte Folksonomie, in der NutzerInnen, Ressourcen und Tags verbunden sind. Das resultierende Netzwerk ermöglicht es, zwischen diesen Entitäten zu navigieren.

In wissenschaftlichen Verschlagwortungssystemen speichern und verwalten BenutzerInnen wissenschaftliche Publikationen. Somit sind solche Systeme ideale Kandidaten für die Untersuchung der Nutzung von Publikationen. In dieser Dissertation untersuchen wir Daten des beliebten Systems BibSonomy und gehen verschiedenen Aspekten der Nutzung von sozialen Verschlagwortungssystemen nach. Darüber hinaus analysieren wir das Potential dieser Daten für Altmetrics, indem wir Korrelationen zwischen der Nutzung und den Zitationen einer Veröffentlichung messen und die Vorhersagekraft bestimmter Nutzungsformen für spätere Zitationen untersuchen.

Wissenschaftliche Verschlagwortungssysteme unterstützen ForscherInnen bei ihrer täglichen Arbeit mit Publikationen und deren Metadaten. Dennoch machen es die schiere Anzahl der verfügbaren Publikationen und deren immer schnelleres Wachstum schwierig, alle relevanten Entwicklungen in einem Forschungsfeld zu verfolgen. Es entsteht ein Problem durch Informationsüberflutung. Hier helfen Empfehlungssysteme, die AnwenderInnen direkt auf bestimmte, für sie relevante Publikationen hinweisen. Dabei kommen personalisierte Ranking-Algorithmen zum Einsatz, die üblicherweise Informationen in Benutzerprofilen ausnutzen, unter anderem die zuvor gespeicherten Ressourcen und die entsprechenden Tags, Informationen über die Ähnlichkeit zwischen einzelnen Entitäten oder die Netzwerkstruktur aller Entitäten (die Folksonomie). In sozialen Verschlagwortungssystemen können solche Systeme der aktiven BenutzerIn sowohl neue Ressourcen empfehlen als auch die Verschlagwortung selbst unterstützen, indem passende Tags vorgeschlagen werden. Wir verwenden das Szenario von Tag-Empfehlungen um das typische Vorgehen bei der Auswertung von Empfehlungssystemen in Folksonomien, bei dem sogenannte Graph-Cores aus Datensätzen gebildet werden, kritisch zu untersuchen. Wir verbessern das Design solcher Experimente durch eine neue, flexiblere Art von Cores, die einen strukturellen Nachteil von Graph-Cores umgehen. Außerdem zeigen wir Gefahren bei der Verwendung von Cores zum Vergleich von Empfehlungsalgorithmen auf. Darüber hinaus beschäftigen wir uns mit der Empfehlung von wissenschaftlicher Literatur. Wir untersuchen und testen verschiedene

---

Möglichkeiten der Integration von Publikationsmetadaten in den populären und vielseitigen Empfehlungsalgorithmus *FolkRank*.




Jedes Web-System muss seinen Dienst rechtskonform und sozialverträglich gestalten. Besonders schwierig ist der Fall von öffentlich sichtbaren Bewertungen, mit denen Produkte von BenutzerInnen beurteilt werden. Dies ist beispielsweise beim Social Peer Review der Fall, bei dem die bewerteten Ressourcen wissenschaftliche Publikationen sind. Wenn die Arbeiten einer ForscherIn (möglicherweise unsachgemäß) kritisiert werden, kann dies Konsequenzen für die Karriere oder für Entscheidungen bezüglich zukünftiger Finanzierung von Projekten haben. Basierend auf Anforderungen, die von deutschem Recht abgeleitet worden sind, beschreiben und diskutieren wir Chancen und Risiken von Web-Systemen, in denen NutzerInnen Publikationen teilen, kommentieren und bewerten.

Insgesamt werden in dieser Dissertation Daten aus dem wissenschaftlichen Publikationszyklus genutzt um Einblicke in Wissenschaftscommunities und die Interaktionen von WissenschaftlerInnen miteinander und mit Literatur zu gewinnen. Wir fokussieren uns besonders auf soziale Verschlagwortungssysteme, in denen die Spuren des Verhaltens von NutzerInnen sichtbar werden und die ein geeignetes Werkzeug darstellen um WissenschaftlerInnen beim Finden und Verwalten von Publikationen zu unterstützen. Das Ziel unserer Beiträge ist es, ForscherInnen in ihrer Arbeit zu unterstützen, als Mitglieder von Forschungscommunities sowie als Produzenten und Konsumenten von wissenschaftlicher Literatur.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	📖 The Scholarly Publication Life Cycle . . . . .	4
1.1.1	Data-driven Studies . . . . .	5
1.1.2	Problem Statements and Studies . . . . .	6
1.2	📌 Scholarly Social Bookmarking Systems . . . . .	10
1.2.1	Application-driven Studies . . . . .	11
1.2.2	Problem Statements and Studies . . . . .	12
1.3	Summary of the Contributions of this Thesis . . . . .	15
<b>2</b>	<b>Foundations and Related Work</b>	<b>17</b>
2.1	Basics and Methodology . . . . .	17
2.1.1	Analyzing Empirical Distributions . . . . .	17
2.1.2	Graphs . . . . .	20
2.2	Analysis of Research Fields and Research Communities . . . . .	26
2.2.1	Analyzing Citations . . . . .	28
2.3	Social Bookmarking Systems . . . . .	31
2.3.1	Folksonomies . . . . .	32
2.3.2	The Social Bookmarking System BibSonomy . . . . .	33
2.3.3	Research Directions on Social Bookmarking . . . . .	37
2.4	Folksonomic Recommender Systems . . . . .	38
2.4.1	Recommendations in Social Bookmarking Systems . . . . .	39
2.4.2	Evaluation of Folksonomic Recommendations . . . . .	40
2.4.3	Folksonomic Recommender Algorithms . . . . .	43
2.5	German Laws with Relevance for Rating and Reviewing . . . . .	44
<b>I</b>	<b>Analyzing Research Communities in Conferences</b>	<b>47</b>
<b>3</b>	<b>📖 Analyzing a Community through its Conferences</b>	<b>49</b>
3.1	Introduction . . . . .	49
3.2	Formal Concept Analysis . . . . .	52
3.3	Related Work . . . . .	54
3.3.1	Formal Concept Analysis . . . . .	54
3.3.2	Formal Concept Analysis of Publications . . . . .	55
3.4	Dataset . . . . .	56
3.4.1	Gathering and Preprocessing . . . . .	56

3.4.2	Notations and Derived Data Structures . . . . .	58
3.5	Analysis . . . . .	59
3.5.1	Conferences . . . . .	59
3.5.2	Authors . . . . .	68
3.6	Conclusion . . . . .	76
3.6.1	Future Research . . . . .	78
<b>4</b>	<b> Analyzing Researchers during a Conference</b>	<b>81</b>
4.1	Introduction . . . . .	81
4.2	Tracking and Supporting Social Interactions at Conferences . . . . .	84
4.3	Related Work . . . . .	87
4.3.1	Conference Applications . . . . .	87
4.3.2	Analysis of Face-to-Face Contacts . . . . .	88
4.4	Dataset . . . . .	89
4.5	Analysis . . . . .	90
4.5.1	Conference . . . . .	90
4.5.2	Workshops . . . . .	92
4.5.3	Peer Groups . . . . .	96
4.6	Conclusion . . . . .	100
4.6.1	Future Research . . . . .	101
<b>II</b>	<b>Analyzing the Usage of Scholarly Social Bookmarking</b>	<b>103</b>
<b>5</b>	<b>  Analyzing Scholarly Publication Management</b>	<b>105</b>
5.1	Introduction . . . . .	105
5.2	The Use Case BibSonomy . . . . .	108
5.2.1	BibSonomy . . . . .	108
5.2.2	Generalizability . . . . .	110
5.3	Related Work . . . . .	112
5.3.1	User Surveys and Post Analysis . . . . .	112
5.3.2	Web Log Mining . . . . .	112
5.3.3	Web Log Mining in Social Bookmarking Systems . . . . .	113
5.4	Dataset . . . . .	115
5.4.1	User and Content Dataset . . . . .	115
5.4.2	Request Log Dataset . . . . .	115
5.5	Analysis . . . . .	115
5.5.1	The Social Aspect . . . . .	116
5.5.2	The Retrieval Aspect . . . . .	121
5.5.3	The Equality Aspect . . . . .	124
5.5.4	The Popularity Aspect . . . . .	130
5.6	Conclusion . . . . .	137
5.6.1	Future Research . . . . .	140

<b>6</b>	<b>📖</b>	<b>Analyzing Publication Usage and Citations</b>	<b>141</b>
6.1		Introduction . . . . .	141
6.2		Alternative Metrics in BibSonomy . . . . .	144
6.2.1		Usage Metrics and Future Citations . . . . .	144
6.2.2		Correlations and Prediction . . . . .	145
6.2.3		Expectations and Limitations . . . . .	147
6.3		Related Work . . . . .	149
6.3.1		Measuring Scholarly Impact in Social Bookmarking Systems . . . . .	151
6.4		Dataset . . . . .	154
6.4.1		BibSonomy . . . . .	154
6.4.2		Microsoft Academic Search . . . . .	155
6.4.3		Matching between BibSonomy and Microsoft Academic Search . . . . .	156
6.4.4		Citation Frequency Distribution . . . . .	157
6.5		Analysis . . . . .	159
6.5.1		Correlations on the Full Corpus . . . . .	159
6.5.2		Correlations for Popular Topics . . . . .	161
6.5.3		Prediction of Future Citations . . . . .	162
6.6		Conclusion . . . . .	165
6.6.1		Future Research . . . . .	166
 <b>III Recommendations and Reviews in Social Bookmarking Systems</b>			<b>167</b>
<b>7</b>	<b>📖</b>	<b>Folksonomic Recommender Evaluation</b>	<b>169</b>
7.1		Introduction . . . . .	169
7.2		Cores of Graphs and Sets . . . . .	172
7.2.1		Generalization . . . . .	173
7.2.2		Examples . . . . .	176
7.2.3		Cores of Folksonomies . . . . .	178
7.3		Related Work . . . . .	183
7.3.1		Graph-Cores . . . . .	184
7.3.2		Evaluation of Recommender Systems . . . . .	184
7.3.3		Sparse Data . . . . .	185
7.3.4		Tag Recommender Systems and their Evaluation . . . . .	186
7.3.5		Cores and Recommender Systems . . . . .	187
7.3.6		Summary . . . . .	188
7.4		Experimental Evaluation . . . . .	189
7.4.1		Datasets . . . . .	189
7.4.2		Evaluation Methodology . . . . .	193
7.5		Results . . . . .	195
7.5.1		Recommendation Performance Depends on Core Type and Level . . . . .	196
7.5.2		Diminished Posts . . . . .	198
7.5.3		Recommender Ranking Correlation . . . . .	199

7.5.4	Exploiting Cores Using <i>LeavePostOut</i>	202
7.5.5	The Most Popular Baseline	203
7.6	Conclusion	204
7.6.1	Lessons Learned	204
7.6.2	Recommendations for Future Tag Recommender Benchmarking Experiments	205
7.6.3	Future Research	205
<b>8</b>	<b>◆ Folksonomic Recommendation of Scholarly Literature</b>	<b>207</b>
8.1	Introduction	207
8.2	<i>FolkRank</i>	210
8.2.1	A Recall of <i>FolkRank</i>	210
8.2.2	The Preference Vector $\vec{p}$	212
8.2.3	Convergence and Variation of <i>FolkRank</i>	212
8.2.4	Extending <i>FolkRank</i>	213
8.3	Related Work	214
8.3.1	Recommending Scholarly Publications	214
8.3.2	Folksonomic Resource Recommendation	215
8.3.3	Exploiting Metadata for Recommendations	217
8.3.4	Improving <i>FolkRank</i> by Including Additional Data	218
8.4	Experimental Evaluation	220
8.4.1	Algorithms	220
8.4.2	Datasets	221
8.4.3	Evaluation Methodology	221
8.5	Results	223
8.5.1	User Similarities	223
8.5.2	<i>FolkRank</i> on an Extended Folksonomy	225
8.5.3	Exploiting Similar Users	227
8.5.4	Exploiting Recent Resources	229
8.6	Conclusion	230
8.6.1	Future Research	232
<b>9</b>	<b>◆ Opportunities and Risks of Online Literature-Reviewing Systems</b>	<b>233</b>
9.1	Introduction	233
9.2	Design Features of Online Rating Systems	236
9.2.1	Rating in Closed User Groups	237
9.2.2	Mode of Rating	238
9.2.3	Aggregated Ratings	239
9.2.4	Ratings in Search Engines	241
9.2.5	Summary	242
9.3	Four Models of Publication Quality Evaluation	242
9.3.1	Classic Peer Review	242
9.3.2	Open Peer Review	243



---

9.3.3	Implicit Ratings . . . . .	243
9.3.4	Social Peer Review . . . . .	244
9.4	Opportunities and Risks of Social Peer-Reviewing Systems . . . . .	245
9.4.1	Opportunities and Risks of Choosing the Peers in Social Peer Review . . . . .	245
9.4.2	Opportunities and Risks of Social Peer Review . . . . .	248
9.5	Realizing a Social Peer-Reviewing System . . . . .	258
9.5.1	A Social Peer-Reviewing System Operated by the Research Community . . . . .	258
9.5.2	Realizing Social Peer Review in a Social Tagging System . . . . .	261
9.6	Conclusion . . . . .	263
9.6.1	Future Research . . . . .	264
<b>10</b>	<b>Conclusion and Outlook</b>	<b>265</b>
10.1	Analysis of Research Fields and Research Communities . . . . .	266
10.2	Social Bookmarking . . . . .	270
10.3	Folksonomic Recommender Systems . . . . .	272
10.4	Further into the Future . . . . .	273
	<b>Appendices</b>	<b>275</b>
	<b>Appendix A</b> References of the Analyzed FCA Publications	<b>277</b>
	<b>Appendix B</b> Correlations of Usage Metrics for Popular Topics in BibSonomy	<b>281</b>
	<b>Appendix C</b> Tag Recommender Results on Different Cores	<b>285</b>
	<b>Appendix D</b> Results from Matrix Theory	<b>289</b>
	<b>Appendix E</b> Personalized Recommender Algorithm Selection	<b>291</b>
	<b>Bibliography</b>	<b>295</b>



## List of Figures

1.1	The two central themes in this thesis. . . . .	3
1.2	The scholarly publication life cycle. . . . .	5
2.1	Relationships between informetrics and four of its subdisciplines. . . .	27
2.2	An example post from BibSonomy. . . . .	33
2.3	Screenshot of BibSonomy. . . . .	34
2.4	Tags of a publication in BibSonomy. . . . .	35
3.1	Citation frequency distributions of the three FCA-minded conferences.	61
3.2	Impact factors of the three FCA-minded conferences over the years. .	63
3.3	Citations from and to papers of the three FCA-minded conferences over the years. . . . .	64
3.4	The concept lattice of authors and their participation at the three FCA-minded conferences. . . . .	67
3.5	A map of co-authorship communities within the participants of the three FCA-minded conferences. . . . .	70
3.6	The concept lattice of influential authors for co-authorship communities.	73
3.7	The concept lattice of influential publications for co-authorship commu- nities. . . . .	74
4.1	A screenshot of the Conferator component TalkRadar. . . . .	86
4.2	Degree and duration distributions at LWA 2010. . . . .	92
4.3	Mined versus workshop induced communities. . . . .	95
4.4	Community roles of LWA 2010 participants. . . . .	99
5.1	Screenshots of BibSonomy's (old) web interface. . . . .	110
5.2	Content visits in BibSonomy over time. . . . .	118
5.3	Retrieval intensity versus self-retrieval. . . . .	119
5.4	Re-visitation behavior of users. . . . .	123
5.5	Requests to different entities over time. . . . .	127
5.6	Transition probabilities between users, tags, and resources in BibSonomy.	129
5.7	Frequency distributions of tags in requests and posts. . . . .	131
5.8	Occurrences of tags in requests and in posts. . . . .	134
5.9	Frequency distributions in requests and posts. . . . .	135
6.1	Citations to publications in BibSonomy. . . . .	155
6.2	Frequency distributions for citations to publications in BibSonomy. . .	157

6.3	Prediction accuracy of usage metrics over future citations (by tag). . .	164
7.1	A folksonomy toy example with various cores. . . . .	179
7.2	Properties of different cores of four folksonomies. . . . .	191
7.3	An example for benchmarking on different cores. . . . .	196
7.4	Precision scores of different recommenders on different cores. . . . .	197
7.5	Benchmarking inconsistency over different cut-levels of precision and recall. . . . .	201
7.6	Tag popularity among posts and users on Delicious and CiteULike. . .	203
8.1	Coverage of withheld items in neighborhoods of similar users. . . . .	223
8.2	<i>FolkRank</i> with high preference values for similar users. . . . .	228
8.3	<i>FolkRank</i> with high preference values for recently posted resources. . .	230
9.1	A resource with two ratings in BibSonomy. . . . .	263
C.1	Recall scores of different recommenders on different cores. . . . .	286
C.2	MAP scores of different recommenders on different cores. . . . .	287

## List of Tables

1.1	Research topics regarding social bookmarking. . . . .	12
3.1	Venues of the three FCA-minded conferences series. . . . .	57
3.2	The history of the three FCA-minded conference series in numbers. . .	60
3.3	Power-law fits to the citation distributions of the three FCA-minded conferences. . . . .	62
3.4	The top five contributing authors to the three FCA-minded conferences.	66
3.5	Top ten rankings of the most influential authors at the three FCA-minded conferences. . . . .	75
3.6	Community roles in the co-authorship graph of the three FCA-minded conference series. . . . .	77
4.1	Contact networks at LWA 2010. . . . .	90
4.2	Distribution of participants over the four workshops of LWA 2010. . .	93
4.3	Community alignment at LWA 2010. . . . .	93
4.4	Actual versus expected number of contacts between workshops. . . . .	94
4.5	Average and median graph centralities per academic status. . . . .	97
5.1	Visiting content in BibSonomy. . . . .	117
5.2	Copying resources. . . . .	121
5.3	Requests to users, tags, and resources. . . . .	126
5.4	Retrieval by tag versus retrieval by search in BibSonomy. . . . .	128
5.5	Correlations and divergence of request and post distributions. . . . .	133
6.1	Correlation between BibSonomy usage metrics and citations (full dataset).	159
6.2	Correlations between usage and current or future citations (full dataset).	160
6.3	Correlations between usage and current or future citations (averaged over popular tags). . . . .	161
6.4	Correlations between usage and future citations (for popular tags). . .	163
7.1	A toy example illustrating different cores. . . . .	176
7.2	An example of a post that is diminished in various cores. . . . .	181
7.3	Sizes of four folksonomy datasets. . . . .	190
7.4	Main cores of four folksonomy datasets. . . . .	193
7.5	Diminished posts in cores of four folksonomy datasets. . . . .	198
7.6	Inconsistent rankings of recommenders over various benchmarking setups.	200
7.7	Cores where a bogus recommender outperforms a baseline. . . . .	202

List of Tables

---

8.1	Properties of two BibSonomy datasets and their cores. . . . .	222
8.2	Coverage of resources among similar users. . . . .	224
8.3	Comparison of different item recommender algorithms on four datasets.	226
8.4	MAP scores of <i>FolkRank</i> with preference manipulation. . . . .	227
8.5	Wins and losses of <i>FolkRank</i> with similar users. . . . .	229
8.6	Wins and losses of <i>FolkRank</i> with recent resources. . . . .	231
B.1	Correlations between usage and current citations by tag stem. . . . .	282
B.2	Correlations between usage and future citations by tag stem. . . . .	283
E.1	Correlation between usage statistics and successful recommendations.	292

## Overview of Author's Contribution

The work presented in this thesis has resulted from several collaborations – mainly with colleagues at the Universities of Kassel and Würzburg. All chapters include previously published content, as described below. References to the respective publications can be found at the end of this overview. Since these publications were collaborations with others, naturally, several of the ideas are the result of team discussions, meetings with my supervisor Prof. Dr. Gerd Stumme as well as with Prof. Dr. Andreas Hotho and Prof. Dr. Alexander Roßnagel, whom I worked with in several research projects.

**Chapter 3** Several parts of the analysis in Chapter 3 have previously been published in [Doerfel et al., 2012b]. The dataset used in the analysis was created jointly by co-author Robert Jäschke and me. Robert Jäschke also computed the co-authorship clustering, visualized in Figure 3.5. It is included in this thesis since it is the basis for the influence analyses in Section 3.5.2. All other analyses were conducted by me and follow my ideas.

**Chapter 4** In Chapter 4, we present results of experiments originally conducted for [Atzmueller et al., 2011] and [Atzmueller et al., 2012b]. While the investigations of roles and academic status (Section 4.5.3) were conducted mainly by me, the discovery of communities was conducted mainly by the co-authors. However, all findings presented here are the result of experiments both extended and revised by me for this thesis. The prototype of Conferator was created by the members of our research group; I shared the responsibility for the TalkRadar and for the user management component in equal parts with Folke Mitzlaff.

**Chapter 5** The chapter is based on the publications [Doerfel et al., 2014c], [Doerfel et al., 2014b], and [Doerfel et al., 2016b]. The main ideas for the analysis come from me. Various ideas have been extended and made more concrete in group discussions with the co-authors. The experiments have been devised and planned by me. Several of the experiments have then been conducted by the co-authors. The dataset has been compiled by co-author Daniel Zoller as part of his master thesis, which I supervised.

**Chapter 6** The results from Chapter 6 have previously been published in [Zoller et al., 2015] and are part of an extended version [Zoller et al., 2016]. The main ideas for the analysis come from me. The experiments have been devised and planned by me. The correlation measurements as well as the prediction experiments have been conducted by co-author Daniel Zoller as part of his master thesis, which I supervised.

**Chapter 7** The results in Chapter 7 have previously been published in [Doerfel and Jäschke, 2013] and [Doerfel et al., 2016a]. I have conducted all experiments and devised the theory, following my own ideas. The diagrams in Figures 7.1, 7.4, and 7.6 have originally been created by Robert Jäschke, based on my experiments.

**Chapter 8** A slight variation of the studies in Chapter 8, using a different version of *FolkRank*, as explained in Section 8.2, has previously been published as [Doerfel et al., 2013a] and before that as [Doerfel et al., 2012a]. The evaluation of *collaborative filtering* and the most popular recommender, serving as baselines for our experiments, has been contributed by Robert Jäschke. All other experiments, as well as the original ideas and settings have been contributed by me.

**Chapter 9** The arguments of Chapter 9 have previously been published in German in [Kartal et al., 2011] and [Doerfel et al., 2013b]. To these publications, I have contributed the technical part, whereas the discussion of judicial aspects is the work of co-author Aliye Kartal-Aydemir. For this thesis, the technical part has been adapted, while the judicial part has been reduced to those aspects that form the basis for the more technical discussion. We point to the respective publications for reference where appropriate.

## References

The following list contains the above mentioned previous publications, which I created during my PhD research, and which have become part of this thesis.

- M. Atzmueller, D. Benz, S. Doerfel, A. Hotho, R. Jäschke, B. E. Macek, F. Mitzlaff, C. Scholz, and G. Stumme. Enhancing social interactions at conferences. *it - Information Technology*, 53(3):101–107, 2011. doi:10.1524/itit.2011.0631.
- M. Atzmueller, S. Doerfel, A. Hotho, F. Mitzlaff, and G. Stumme. Face-to-face contacts at a conference: Dynamics of communities and roles. In M. Atzmueller, A. Chin, D. Helic, and A. Hotho, editors, *Modeling and Mining Ubiquitous Social Media - International Workshops MSM 2011, Boston, MA, USA, October 9, 2011, and MUSE 2011, Athens, Greece, September 5, 2011, Revised Selected Papers*, volume 7472 of *Lecture Notes in Computer Science*, pages 21–39. Springer Berlin Heidelberg, Heidelberg, Germany, 2012. ISBN 978-3-642-33683-6. doi:10.1007/978-3-642-33684-3.2.
- S. Doerfel and R. Jäschke. An analysis of tag-recommender evaluation procedures. In *Proceedings of the 7th Conference on Recommender systems, RecSys '13*, pages 343–346, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2409-0. doi:10.1145/2507157.2507222.
- S. Doerfel, R. Jäschke, A. Hotho, and G. Stumme. Leveraging publication metadata and social data into folkRank for scientific publication recommendation. In *Proceedings of*



- the 4th ACM RecSys workshop on Recommender systems and the social web, pages 9–16, New York, NY, USA, 2012a. ACM. doi:10.1145/2365934.2365937.
- S. Doerfel, R. Jäschke, and G. Stumme. Publication analysis of the formal concept analysis community. In F. Domenach, D. Ignatov, and J. Poelmans, editors, *Formal Concept Analysis – 10th International Conference, ICFCA 2012, Leuven, Belgium, May 7-10, 2012. Proceedings*, volume 7278 of *Lecture Notes in Artificial Intelligence*, pages 77–95, Berlin/Heidelberg, May 2012b. Springer. ISBN 978-3-642-29892-9. doi:10.1007/978-3-642-29892-9\_12.
- S. Doerfel, A. Hotho, A. Kartal-Aydemir, A. Roßnagel, and G. Stumme. Empfehlungssysteme für wissenschaftliche Publikationen. In *Informationelle Selbstbestimmung im Web 2.0*, pages 113–148. Springer Berlin Heidelberg, 2013a. ISBN 978-3-642-38055-6. doi:10.1007/978-3-642-38056-3\_6.
- S. Doerfel, A. Hotho, A. Kartal-Aydemir, A. Roßnagel, and G. Stumme. *Informationelle Selbstbestimmung im Web 2.0 – Chancen und Risiken sozialer Verschlagwortungssysteme*. Xpert.press. Springer Berlin Heidelberg, 2013b. ISBN 978-3-642-38055-6. doi:10.1007/978-3-642-38056-3.
- S. Doerfel, D. Zoller, P. Singer, T. Niebler, A. Hotho, and M. Strohmaier. Evaluating assumptions about social tagging: A study of user behavior in BibSonomy. In T. Seidl, M. Hassani, and C. Beecks, editors, *Proceedings of the 16th LWA Workshops: KDML, IR and FGWM, Aachen, Germany, September 8-10, 2014*. CEUR-WS.org, 2014a. URL <http://ceur-ws.org/Vol1-1226/paper06.pdf>.
- S. Doerfel, D. Zoller, P. Singer, T. Niebler, A. Hotho, and M. Strohmaier. How social is social tagging? In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion, WWW Companion '14*, pages 251–252, Republic and Canton of Geneva, Switzerland, 2014b. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-2745-9. doi:10.1145/2567948.2577301.
- S. Doerfel, R. Jäschke, and G. Stumme. The role of cores in recommender benchmarking for social bookmarking systems. *ACM Transactions on Intelligent Systems and Technology*, 7(3):40:1–40:33, February 2016a. doi:10.1145/2700485.
- S. Doerfel, D. Zoller, P. Singer, T. Niebler, A. Hotho, and M. Strohmaier. What users actually do in a social tagging system: A study of user behavior in BibSonomy. *ACM Transactions on the Web*, 10(2):14:1–14:32, May 2016b. doi:10.1145/2896821.
- A. Kartal, S. Doerfel, A. Roßnagel, and G. Stumme. Privatsphären- und Datenschutz in Community-Plattformen: Gestaltung von Online-Bewertungsportalen. In H.-U. Heiß, P. Pepper, H. Schlingloff, and J. Schneider, editors, *Informatik 2011 - Informatik schafft Communities - Proceedings der 41. GI-Jahrestagung*, volume 192 of *Lecture*

*Notes in Informatics*, page 412. Gesellschaft für Informatik e.V. (GI), Bonner Köllen Verlag, 10 2011. URL <http://www.informatik2011.de/541.html>.

D. Zoller, S. Doerfel, R. Jäschke, G. Stumme, and A. Hotho. On publication usage in a social bookmarking system. In *Proceedings of the ACM Web Science Conference, WebSci '15*, pages 67:1–67:2, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3672-7. doi:10.1145/2786451.2786927.

D. Zoller, S. Doerfel, R. Jäschke, G. Stumme, and A. Hotho. Posted, visited, exported: Altmetrics in the social tagging system BibSonomy. *Journal of Informetrics*, 10(3): 732–749, 2016. ISSN 1751-1577. doi:10.1016/j.joi.2016.03.005.

# Chapter 1

## Introduction

In this thesis, we present studies analyzing research and researchers and their interactions recorded at conferences or with web-based tools, particularly with the social bookmarking system BibSonomy. The overall goal of these investigations is to support researchers in their work with scholarly literature. In the following, we describe the broader vision to which these studies contribute, as well as the two main themes of the thesis: the utilization of data from the publication life cycle and scholarly social bookmarking.

*“Which articles should I read? Who are the key players in my area of research? Who should I approach at the next conference? Which publication will be important next year? How do I find important content and related work?”*

Researchers must ask themselves these and similar questions time and again to keep up-to-date in their respective fields of research, to adjust their own research directions, to collect meaningful related work sections for their own publications, and to choose partners for collaboration. Finding good answers is particularly hard for newcomers – those who just begin their career as a researcher or those who plan to approach a new field of research broadening their area of interests – as they have to start from scratch. It can, however, even be challenging for already established researchers due to the phenomenon of information overload, that is, the overabundance of scholarly literature. Respective complaints have been voiced already back in the early days of science, even before the first scientific journals were published. For instance, Price [1963] quoted a statement from 1613 by scholar Barnaby Rich about the number of books which “overcharge the world that it is not able to digest the abundance of idle matter that is every day hatched and brought forth into the world.” Today it is well-known, that the number of available scholarly publications (and of researchers) does not only grow, but it grows ever faster. Price [1963] discussed characteristic numbers of science, like the number of scholarly publications (in a research area, or in scientific journals) or the number of researchers, and he showed that they mostly grow exponentially. Depending on the discussed index, he observed doubling rates of 10 to 15 years.<sup>1</sup>

To familiarize themselves with the state of the art in a particular research area, or to keep up-to-date with it, researchers can be supported by showing them analyses of

---

<sup>1</sup>More recently, by counting publications in the reference sections of articles listed in the Web of Science, Bornmann and Mutz [2015] estimated an enormous growth rate of 8.9% per year, implying a doubling of the number of scholarly publications roughly every eight years.

that area. Especially in the face of the exponentially growing publication output, such analyses are useful to grasp and to explore the respective research area from different angles. Thus, it suggests itself to offer visualizations of a research community with its key players, to highlight particularly important work and authors (according to a variety of suitable measures), or to display influences in communities, for example, publications that are particularly relevant for a subcommunity. Analyses of research fields or of research communities belong to the field of scientometrics and they are published for various research areas, for instance, in the *Scientometrics Journal*. They usually rely on the publications in the respective fields or their metadata.

Next to these scientific analyses, tools have been created for researchers, that can assist them with various tasks: Scholarly publication management systems, like BibSonomy,<sup>2</sup> CiteULike,<sup>3</sup> and Mendeley,<sup>4</sup> allow researchers to collect, organize, and share publication metadata. Bibliographic databases, like the Web of Science<sup>5</sup> or Scopus,<sup>6</sup> maintain selected sets of research articles and make them available for subscribers. Scholarly search engines, like Google Scholar<sup>7</sup> and Microsoft Academic Search,<sup>8</sup> allow querying for publications from a corpus of publications gathered through crawling the web. Similarly, search engines have been included in the previously mentioned publication management systems, relying on the collaboratively collected content. Altmetrics and social peer review have been established to identify particularly relevant publications (either identified through their usage in social media or by expert opinions). Finally, scholarly publication recommender systems compute personalized lists of publications that are likely to be relevant to the active user.

In this thesis, we contribute to both forms of supporting researchers: We present new analyses of research communities and we scrutinize and extend tools for scholars, such as social bookmarking systems and their extensions with recommendation and reviewing features, as well as a guidance system for scholarly conferences, to improve their usefulness for their audience. In the following, we present an overview of this thesis and address the open research questions. Overall, our contributions follow two main themes:

**Theme 1:** We generate new analyses from data of different phases of the scholarly publication life cycle, covering the creation, dissemination, usage, and citation phase of a publication (see Section 1.1). These analyses are helpful for researchers, who can use them to better understand their communities, and for operators of scholarly web tools, who can improve their systems (based on the analyses) for the benefit of their target audience.

---

<sup>2</sup><http://www.bibsonomy.org/>

<sup>3</sup><http://www.citeulike.org/>

<sup>4</sup><http://www.mendeley.com/>

<sup>5</sup><http://ipsience.thomsonreuters.com/product/web-of-science/>

<sup>6</sup><http://www.scopus.com/>

<sup>7</sup><https://scholar.google.com/>

<sup>8</sup><http://academic.research.microsoft.com/>

---

◆ **Theme 2:** We study various aspects of (scholarly) social bookmarking systems (see Section 1.2) to determine how such a system is used and how it can be improved and extended to allow more efficient work with literature.

With Theme 1, we contribute to the field of scientometrics, relying on the traditional data sources, data from the creation and the citation phase of a scholarly publication life cycle, as well as on data from the usage and from the previously rarely investigated dissemination phase. With the former, we add to the altmetrics discourse, which is concerned with alternatives to citation-based analyses. With the latter, we analyze researchers at a conference, thus, an event where publications are presented and discussed. The second theme focuses on scholarly social bookmarking systems and thus on data from the usage phase of the publication life cycle. In such systems, users can store, organize, and retrieve publications, and they can discover publications that other users have stored. Thus, these systems support users in the handling of their literature collections. Additionally, that data can be used to compute recommendations and thus support researchers during the creation of new publications, for instance, by suggesting articles relevant to their lines of research.

In the following, we describe these two themes – the publication life cycle and social bookmarking systems – in more detail, and we motivate the research problems that we tackle in this thesis. The chapters in this thesis contribute either to one or to both of these goals, and we use the book symbol (📖) and the tag symbol<sup>9</sup> (◆) to mark the chapters as belonging to one (or both) of these two themes. Figure 1.1 shows the chapters of this thesis and the themes to which each chapter belongs. Part I contributes to the first theme, in Part II the themes overlap, and Part III contributes to the second theme.

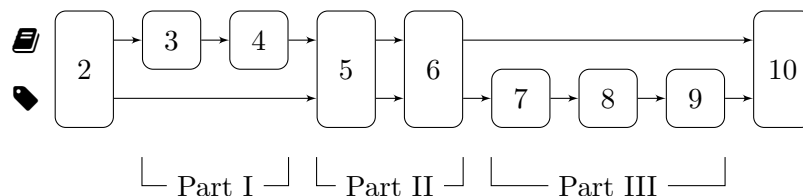


Figure 1.1: The two central themes in this thesis: The diagram shows the parts and chapters of this thesis and to which of the thesis’s two main themes each of the chapters belongs. Chapters 3 through 6 belong to the scholarly publication life cycle theme (📖), and Chapters 5 through 9 to the social bookmarking systems theme (◆). Chapter 2 recalls methodology that is relevant for both contexts, and Chapter 10 contains an outlook.

---

<sup>9</sup>Social bookmarking systems are often also called social tagging systems or collaborative tagging systems, emphasizing the central feature of annotating resources with tags. We will use these terms interchangeably in this thesis.

Before we begin with our studies in Part I, in Chapter 2, we recall various notions and methodology that will be used in several studies throughout this thesis. Moreover, we describe the framework of informetrics and where our studies fit into that context. We also recall the most relevant aspects of social bookmarking and of recommender systems.<sup>10</sup> The thesis is concluded with perspectives on future developments and opportunities for further research in Chapter 10.

## 1.1 The Scholarly Publication Life Cycle

The contributions in Part I and Part II of this thesis are aligned along the life cycle of a scholarly publication. Here, we presume a relatively simple cycle, comprising the four phases *creation*, *dissemination*, *usage*, and *citation*, as depicted in Figure 1.2:<sup>11</sup> This model is an adaptation of the information cycle described by Nagelschmidt [2010] which contains the steps *generation and publication*, *establishing and providing*, *acquisition and management*, and *adoption and processing*. The scholarly publication life cycle describes these four steps from the point of view of a single publication: A publication is *created* by its authors, who conduct the research, produce the text, and insert references to other publications. After the resulting manuscript has been accepted at some publication venue, the creation phase ends with the actual publication. The publication is then *disseminated*, for instance, by distributing it in a journal issue or by presenting it in a talk at a conference. Afterwards, the publication can be *used* – acquired, borrowed, viewed, downloaded, read, stored, tagged, and so on. Eventually, if it is relevant to new research, the publication is *cited*. Through their references, the life cycles of different publications are connected, the citation phase of one publication is the creation phase of another. In Figure 1.2, publication *B* is used and then cited in *A*, and in turn *A* is cited in *C* – leading to overlapping life cycles of the three publications.

Scientometrics has traditionally utilized the publication metadata that is produced in the creation phase (e.g., analyzing co-authorships, keywords, or titles) and the citation phase (e.g., comparing publications and authors by measures computed from the citations they received). However, more and more of the research processes and workflows have become observable: Researchers manage their publications online instead of in some private (offline) collection, for example, in dedicated bookmarking systems; publications are downloaded and viewed in digital libraries; and researchers discuss and even rate articles on respective web platforms. Thus, the usage phase of a publication’s life cycle also produces data, which can be collected and can serve as a source for research on researchers and scholarly publications. Finally, new (experimental) RFID technology enables tracking researchers during a conference. This

---

<sup>10</sup>Since in later chapters, we refer explicitly to specific sections of Chapter 2 when the respective concepts are used, readers who are familiar with the topics of this thesis, might go directly to the studies beginning with Chapter 3, and come back to Chapter 2 when needed.

<sup>11</sup>We will also use the visualization from Figure 1.2 in each Chapter of the thesis, to highlight the phase of the publication life cycle that plays the main role in that chapter.

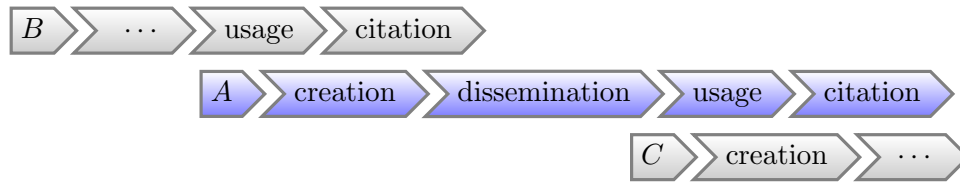


Figure 1.2: The scholarly publication life cycle, covering the four phases *creation*, *dissemination*, *usage*, *citation*. The figure shows the life cycle for publication *A*, together with that of a publication *B*, which *A* might cite, and with that of a publication *C*, which might cite *A*.

allows the study of interaction networks between researchers during such events, which belong to the dissemination phase of the publication life cycle.

### 1.1.1 Data-driven Studies

The analyses in Part I and Part II are data-driven by the scholarly publication life cycle in the sense that we exploit data that is produced in one of the cycle’s phases. We use that data to analyze the community (the group of researchers) which produced it, or the system from which the data has been collected. More particularly, we make use of the following data, obtained from the four phases of the publication life cycle:

**Creation: Publication Metadata.** A publication is written and included in a journal, in conference/workshop proceedings, or at least made available in some repository, for example, as a preprint or technical report. In that process, metadata about that publication (its title, authors, publication year, venue, references, etc.) is made available in print or online. This kind of data is the classical source for the analysis of a research area.

**Dissemination: Contacts and Locations of Researchers at a Conference.** Usually, a contribution to a conference or workshop is not only published in proceedings, but also presented at the respective event. For such occasions, researchers come together and interact with each other – to discuss results or to start or continue collaboration. Contact and location data captures the information of where a participant of a conference is and who he meets. We will make use of data that has been gathered from RFID sensors worn by participants of a conference.

**Usage: Publication Usage Data in a Publication Bookmarking System.** A scholarly publication can attract the interest of the scientific community.<sup>12</sup> Researchers

<sup>12</sup>Usually that happens after it has actually been published. However, it has become common practice in some areas of research to make manuscripts publicly available through preprint servers like the arXiv (<http://arxiv.org/>) before an “official” publication in a peer-reviewed venue.

collect and cite interesting publications and for that purpose use software to manage their collections. A dataset from the social publication management system BibSonomy allows us to investigate how researchers make use of such software and to what extent they profit from the collections of others.


**Citation: Cited-by Relations.** A publication’s value for one’s own work is usually acknowledged by citing it appropriately. The citations to a publication can be counted from the reference sections of other publications. Thus, the “cited-by” relation is created during a publication’s citation phase – or respectively during the creation phase of the citing publications. Citation data can be found in large publication corpuses, in our case that of the scholarly search engine Microsoft Academic Search.

### 1.1.2 Problem Statements and Studies

Following the phases of the publication life cycle, we attend to different research questions in the investigation of research communities. Among others, we employ methods that have not previously been used in the respective field, to provide new insights that facilitate an overview about a research community, and a better understanding of the users of a scholarly literature tool. While in each of the Chapters 3 through 6, we employ a specific use case – a research community or a scholarly web system – for demonstrating the methods, all of them are generalizable to other communities or similar systems. As a side effect, we provide detailed studies of those communities that serve as use case.


#### **Part I: Analyzing Research Communities in Conferences**

The first part of the thesis focuses on the analysis of research communities, relying on data from the first two phases of the scholarly publication life cycle, particularly data that is gathered from research conferences of the respective community.

 **Analyzing a Community through its Conferences (Chapter 3).** An overview about a research community, its key figures, its most relevant publications, and composition of various subcommunities can be a great help for researchers who want to join the community and even for researchers who are already a part of it (due to the issue of exponential growth, making it ever harder to keep up with the state of the art). Such an overview can be the starting point for the selection of literature-to-read or of researchers to talk to. Many scientometric analyses of research communities rely on the output of journal publications listed in the Web of Science. Yet, this ignores a large number of publications – particularly conference contributions, which however are usually a much faster means of contributing to the research discourse in a community. Therefore, we approach the task of describing a community through its conferences. Moreover, instead of focusing purely on statistical analyses, like it is often encountered in scientometrics (e.g., productivity per year, country, or researcher), we



additionally use methodology from formal concept analysis (see Section 3.2). As a complementary approach to scientometric analyses, this gives us the means to produce easy-to-read, concise visualizations of relations between conferences (visiting patterns) and between publications and authors (influences on particular subcommunities). Comparing conferences in this way can help those who steer the respective events, showing their own series in the context of the others, as well as those who want to select a conference for their next publication. Analyses of the influences within the community and particularly of influences on various subcommunities inform on the history of the subject and are interesting for those who want to work on the same topics as one of the subcommunities and who are looking for relevant work based on its impact on this particular group of authors rather than on overall impact (like in the traditional comparing of impact by counting citations). Chapter 3 is the first chapter of our journey through the publication life cycle, relying on data from the creation phase of publications.

 **Analyzing Researchers during a Conference (Chapter 4).** We zoom in from the comparison of several conferences series of the same community to one single conference. Although it is assumed that conferences are valuable occasions for researchers to come together and exchange ideas, the actual interactions of participants of such events have rarely been studied. For organizers of such an event it is difficult to assess whether their goals, like supporting exchange and dialogue between different members of different sub-areas or integrating newcomers into a the community, have been reached during the conference. Meeting these goals is a decisive factor for the success of a research conference.



Therefore, in Chapter 4, we present analyses that allow insights into the interactions between members of different subcommunities and on the standing of newcomers (students and PhD students) compared to established researchers (post-doctoral researchers and professors). Such analyses benefit conference organizers by giving an impression on the extent to which subcommunities mix, and by helping observe the roles of different groups of participants. They are also helpful for participants, who can judge their own role in comparison to those of others. Moreover, we present and describe a new tool, Conferator, which serves as a conference guidance system for participants, facilitating networking with colleagues and managing the conference schedule.

Whereas in the previous chapter, we analyzed a community through its work, that is, its output of scholarly publications, in Chapter 4, we focus on data of personal interactions between members of the community, thus data of the dissemination phase of the publication life cycle.

## **Part II: Analyzing the Usage of Scholarly Social Bookmarking**



The second part of the thesis belongs to both themes of this thesis. After publications have been disseminated, they can be used by researchers – they can be downloaded,

saved, marked as relevant, shared, and, of course, read. Scholarly social publication management systems, like BibSonomy, CiteULike, or Mendeley, are dedicated tools that assist researchers in collecting and organizing the publications they use. While there are various systems in which publications can be used (e.g., digital libraries or publisher catalogs), bookmarking systems are of particular interest, since, next to the fact that they document how users create and use their own collections, they also have a collaborative aspect that makes them part of the family of social software. Users cannot only interact with their own collection, but also browse and inspect the collections of others. Users might collect publications for purely selfish reasons, yet others can still profit from these activities (as we will show in Chapter 5). The resulting usage data (in our use case from the social bookmarking system BibSonomy) allows us to analyze four aspects of social tagging that have been discussed controversially in the literature (Chapter 5). A fifth aspect – the altmetrics aspect – is specific to *scholarly* social tagging systems and therefore discussed in its own chapter (Chapter 6).

  **Analyzing Scholarly Publication Management (Chapter 5).** Already early after the first social bookmarking systems had been created, their central feature – publicly organizing resources using tags – has received praise, and much research has focused on the data structures emerging from tag assignments, so-called folksonomies (see Section 2.3.1). However, only little is known about what users actually do in such systems and to what extent they make use of the possibilities offered to them. A better understanding of user behavior is relevant for those who operate such systems. It allows them to evaluate whether the features they offer are successful, and it can be the starting point for improving their system, for instance, to reduce the effort it takes to find and manage relevant publications. Thus, albeit indirectly, an analysis of user behavior in a social bookmarking system is also beneficial for the system’s users – in the case of scholarly bookmarking, the researchers who organize and share their publications.

In Chapter 5, we employ the use case of the scholarly bookmarking system BibSonomy and analyze four aspects of its usage: socialness, personal management, the importance of users, tags, and resources for navigation, and popularity. The four aspects have been discussed controversially in the literature. For example, while social tagging systems are called social, it is unclear to what extent users actually use these systems to exchange content and to what extent they just enjoy the personal management of their own collection. While, for example, Weinberger [2005] called the social aspect “highly useful”, Vander Wal [2005] suggested that the main reason for using tagging systems was personal management. The interest, users show for the content and tags of other users (i.e., their use of the social aspect of tagging) is revealed in their requests to the system. Through analyzing these requests, found in the server log files of BibSonomy, we are able to capture the traces of actual human behavior and to observe what users actually do (as opposed to what they think or claim to do – like in

studies relying on questionnaires). For each of the above mentioned four aspects, we discuss the observed evidence from the log files.

  **Analyzing Publication Usage and Citations (Chapter 6).** With the advent of the Web 2.0, the usage of publications in social media or online managers has become observable, and the question of what we can learn from comparing publications by their usage intensity, has arisen. Traditional scientometrics have largely ignored these activities as data source, yet, at the latest with the coinage of the term *altmetrics*, the study of the usage of publications in social media has emerged as an area of research. One of the central questions in this field is that of the relationship between usage-based and citation-based measures, and correlations of both have been investigated in a variety of studies (see Section 5.3). With Chapter 6, we add to the altmetrics discourse by addressing the altmetric aspect in BibSonomy (for the first time). In contrast to the four aspects in the previous chapter, this fifth aspect of social bookmarking is specific to *scholarly* applications.

Social bookmarking systems have been studied as sources for altmetrics mainly with regard to the number of posts, while indicators like exports, views, or all requests to a publication in general, have been ignored. However, such measures are available in publication management systems as well, and in Chapter 5, we will see that post and request counts are not strongly correlated, suggesting that they will yield rather diverse metrics. Therefore, in Chapter 6, we study how both relate to citation-based impact. Moreover, most of the previous studies regarding altmetrics in publication management systems have focused on Mendeley and on journal publications listed in the Web of Science. Thus, the situation in other systems and particularly with other publication types is unclear. However, ignoring conference proceedings does not accurately reflect the publication reality in some research disciplines, like computer science.<sup>13</sup> To account for that, in our study, we do not discriminate by publication type and allow all publications that users have posted to BibSonomy.

Finally, instead of only measuring correlations, we attempt to predict citation impact. Here, we only use the data available in a social bookmarking system,<sup>14</sup> thus a realistic setting for those who operate such a system. With the studies in this chapter, we conclude our investigations along the publication life cycle by comparing data from the usage phase to data from the citation phase of publications.

---

<sup>13</sup>In her summary of comments from various researchers on the publication cultures in their respective disciplines, Schuh [2009] explained how diverse the preference of particular publication types is. For instance, in literature studies, contributions to proceedings are held in higher regard than journal articles. Overall in the humanities, monographs seem to be the most important means of publishing. Increasingly popular become pre-prints, in which new research can be made publicly available much faster than by going through publishing processes of journals. Thus, also citations can occur much faster when they appear in pre-prints.

<sup>14</sup>In particular, we refrain from using external classifications of publications into research fields.

## 1.2 Scholarly Social Bookmarking Systems

The second main theme of this thesis is the analysis and improvement of scholarly social bookmarking systems. While such systems have assisted researchers in their work with literature for years, there are still several aspects that have not yet been fully investigated: (i) While much has been speculated about the usefulness of this form of resource management, little has been investigated what users actually do in such a system and how the central features of social bookmarking are used to retrieve resources. Moreover, the rising interest in altmetrics has prompted the question of what can be concluded from the usage of publications in these systems [Priem et al., 2010]. (ii) Today’s social bookmarking systems have long been developed beyond the plain tagging functionality to include many further features, like the Web 2.0-typical recommender systems and reviewing options. Open lines of research are the improvement of these features as well as aspects of social and legal compatibility. In the following, we describe the scenario of scholarly social tagging, entertaining both the central functionality (tagging and retrieving) as well as the extended features (recommendations and reviews). Afterwards, in Section 1.2.2, we describe the road map for our research on scholarly bookmarking in this thesis and the individual research questions of each chapter.

Scholarly publication management tools support researchers in their work with literature. Through tagging, users can organize their collections of literature using freely chosen keywords, called *tags*. Using these tags, they can retrieve the stored resources later on. Usually, the resources together with the annotations are publicly visible, thus researchers can browse the collections of others and they can make use of the system’s literature corpus that arises from the individual collections of all users. Tagging systems and their data have proven to be valuable sources for researchers both as tool and as subject of studies (for examples see Section 2.3). Furthermore, users of tagging systems have expressed their desire for tagging in other systems as well [Noy et al., 2008], and so the practice of tagging has found its way as a secondary feature into many web systems, for instance, tagging of products in online shopping portals, of articles in blogs, and of messages in microblogging systems and social networks (hashtags). Therefore, our contributions to the field of tagging can be expected to be relevant for such systems (scholarly or non-scholarly) as well.

Naturally, social tagging systems have been extended beyond the plain tagging functionality, for instance, by recommendation and reviewing features. Both recommending and (post-publication-)reviewing of publications are activities from the usage phase of publications.<sup>15</sup> Recommender systems assist users during the tagging process

---

<sup>15</sup>Scholarly recommender systems have been proposed for and relying on other phases of the scholarly publication life cycle: For example, Heck et al. [2011] used data from the social bookmarking system CiteULike – data from the creation and from the usage phase of the publication life cycle – to recommend collaborators (selected from the authors of publications in that system). Moreover, recommendations are not only being offered to authors but also to editors or program chairs who decide over a manuscript’s publication, by proposing appropriate reviewers: Liu et al. [2014]

(by suggesting tag candidates) or by directly presenting resources that they might like. Usually, they produce ranked lists of personalized recommendations, relying on algorithms that utilize a user's previous interests and activities. In the case of scholarly tagging systems, recommender systems assist researchers by suggesting tags for publications or by pointing them towards those publications that are likely to be relevant to their research, thus sparing them (or at least alleviating) the work of going through the vast body of available literature manually. The development and enhancement of folksonomic recommender systems is still an active field of research, as better algorithms can improve the experience for users, who are guided to more relevant publications or who receive better suggestions during the tag selection process. In this thesis, we critically investigate the evaluation framework for folksonomic recommender algorithms, and we conduct studies on various extensions of the popular algorithm *FolkRank*.

While recommendations focus on the task of automatically providing personalized suggestions, optimized towards relevance for a specific user, reviewing is rather a form of quality control to which the users contribute by providing their personal opinion. In scholarly systems, researchers provide their opinion on scholarly publications – a process called social peer review. Scholarly publication tagging systems suggest themselves as platforms for social peer review since, thus, the process of writing reviews is integrated into the same context as the personal management of publications. Readers can let others know about their experience with a publication: They may praise it (by assigning it a high rating or by positive discussions) and, thus, contribute to its attractiveness for others. However, they can also warn their colleagues or point to mistakes. Thus, they aid researchers who might use and cite a publication in their work. Aggregating all reviews for the same publication yields an overall quality assessment. The reviewing feature can be used to highlight excellent publications or to denote redundant or faulty ones, thus helping others find the good ones and avoid the bad ones. In the thesis, we discuss opportunities and risks of such a feature in the light of social compatibility.

### 1.2.1 Application-driven Studies

The studies on social bookmarking in this thesis are driven by application. We focus on the popular scholarly social bookmarking system BibSonomy, which is operated and developed by our research group<sup>16</sup> for two purposes: (i) It provides support to

---

suggested reviewers for submitted manuscripts and optimized a list of recommended reviewers towards authority, expertise, and diversity. Candidates (for reviewing) were described by their co-authors and the content of their publications (i.e., data from the creation phase of the publication life cycle). Utility during the dissemination phase is provided, for example, by the *acquaintomatic* feature [Atzmueller et al., 2012a] in the system Conferator (introduced in Chapter 4), where conference participants are recommended to each other as conversational partners. Another example is given by Lee and Brusilovsky [2014], who described and evaluated talk recommenders for conference participants based on their system Conference Navigator.

<sup>16</sup>Both is conducted in cooperation with groups at the Universities of Würzburg and Hanover.

Table 1.1: Research topics of the thesis’s social bookmarking theme. In Parts II and III we cover the core functionality of tagging as well as recommendations and reviews in social bookmarking systems. We conduct studies on their usage or their algorithmic improvements and their social and legal compatibility depending on the availability of previous work and of suitable data. We point to literature for topics that have already been covered elsewhere.

	usage / algorithms	social and legal compatibility
tagging (core features)	Chapters 5 and 6	e.g., Krause et al. [2010]
recommendations	Chapters 7 and 8	e.g., Doerfel et al. [2013a]
rating and reviewing	future work	Chapter 9

researchers all over the world for managing and finding literature, and, (ii) through its usage, BibSonomy produces datasets containing both the resources and annotations that form the users’ collections (the posts), as well as traces of the users’ activities in the system (the request logs). These datasets allow us to study what users actually do in the tagging system as well as to develop algorithms (e.g., for recommendation) by using the datasets for training and testing. While datasets containing the public posts have been made available from BibSonomy and other systems (e.g., Delicious or CiteULike), previously no data on the actual usage of a tagging system has been made available. Thus, our BibSonomy dataset, which contains next to the posts also all user requests to the system, is a unique novel opportunity to study actual usage beyond the collecting and tagging itself.

Note that, although several studies rely solely on BibSonomy data, the procedure of the analyses could easily be applied to other systems. Whether or not other systems would yield similar results is a matter of speculation – the unavailability of suitable data is the bottleneck.

## 1.2.2 Problem Statements and Studies

In our studies of social bookmarking systems, we follow two lines of research: algorithmic improvement, as well as alignment with social and legal norms. Some of these aspects have already been the subject of previous research. Hence, we focus on research questions that have rarely been investigated or where we can demonstrate improvements on the state of the art in our studies. Table 1.1 summarizes the research areas to which each chapter contributes (and points to literature for the other cases, see below).

## **Part II: Analyzing the Usage of Scholarly Social Bookmarking**


Our studies related to the social bookmarking theme of the thesis begin in Part II.<sup>17</sup> Since the two themes of the thesis overlap particularly in this part, we have already mentioned the studies of Chapters 5 and 6 on the basic features in the previous section. In both chapters, we provide analyses on various aspects of the usage of social bookmarking. We address and question typical beliefs about social tagging and we assess the potential of usage data in such a system as a source for altmetrics. Legal aspects of the central features in social bookmarking systems are not part of the thesis as they have already previously been debated by Krause et al. [2010], especially regarding the requirement of data protection. They reviewed design choices and data collection during registration, search, and posting.

## **Part III: Recommendations and Reviews in Social Bookmarking Systems**

In the third part of the thesis, we move on to common extended features of bookmarking systems, namely recommendation and reviewing functionality. In contrast to the basic features which enable the management of publications, these extensions are more proactive approaches of supporting researchers: recommendations point researchers directly to relevant literature and reviewing allows readers to express their opinion and to take influence on the success of a publication.

Recommender systems are the topic of Chapters 7 and 8. The alignment with social and legal norms of recommender systems has already been discussed in [Doerfel et al., 2013a], particularly regarding the opportunity of misusing the intransparency of recommendation algorithms to manipulate the selection of recommended products to suit the needs of the provider rather than optimizing the recommendations towards the interests of the consumer. Therefore, here, we focus on the algorithmic aspect of recommendations, by discussing benchmark settings to compare algorithms and by improving a popular algorithm, *FolkRank*, for publication recommendation.

Online reviewing, and particularly social peer review of scholarly publications is addressed in Chapter 9. We have implemented a reviewing and rating feature in BibSonomy. However, since the feature is relatively young, only few publications have been rated by many users. Therefore, we leave quantitative analyses or the exploitation of ratings for rankings and recommendations for future work. Instead, we rather discuss the feature itself in the context of legal and social compatibility (Chapter 9).

 **Folksonomic Recommender Evaluation (Chapter 7).** Recommender systems in tagging systems have been investigated frequently in the past. The most common approach for their evaluation is the use of a historic dataset to benchmark new algorithms or novel modifications of existing ones. For that scenario, a particular form

---

<sup>17</sup>The system Conferator, introduced in Chapter 4 is integrated with BibSonomy. However, it is itself its own system and the analyses in Chapter 4 focus rather on the interactions of researchers than on their bookmarking activities.

of preprocessing – restricting the original dataset to a so-called  $p$ -core – has become common practice. Albeit used in various recommender benchmarks, the influence of that preprocessing has never been analyzed and the chosen  $p$ -cores have often been selected without analytic justification. In Chapter 7, we question this practice and show that using cores indeed introduces several problems. We will use datasets from various tagging systems to critically inspect the consequences of choosing such cores, point to various pitfalls and show that using cores carelessly yields unstable benchmarking results. Some of these issues can be overcome using a new type of core, a *set-core*, which we introduce. Other issues remain and cannot easily be fixed. For these cases, we present a list of recommendations for setting up a benchmarking scenario that will yield valid results.

◆ **Folksonomic Recommendation of Scholarly Literature (Chapter 8).** As already mentioned above, recommending scholarly publications is a formidable means to mitigate the information overload problem that researchers face in their work with literature. Since scholarly social bookmarking systems have a corpus of publications available, it suggests itself to compute recommendations from that corpus for the system’s users. Moreover, it is another way of sparking the interest of users in the content of others (in Chapter 5, we will see that visiting the content of others accounts for a significant part of the overall interaction with BibSonomy, but it is not the dominant use case).

Usually in social bookmarking systems, recommendation algorithms tap into the folksonomy data structure, that is, the connections between users, tags, and resources that are contributed by the users through annotating resources (tag assignments). However, the folksonomy structure is not the only available data in such systems. There is also a limited amount of publication metadata, as well as social connections between the users. In Chapter 8, we explore how such data can be leveraged into the well-known and established folksonomic recommendation algorithm *FolkRank*, to recommend scholarly publications. We follow the results from Chapter 7 to conduct a benchmarking of several *FolkRank*-based recommendation strategies, and we compare them to several baselines.

◆ **Opportunities and Risks of Online Literature-Reviewing Systems (Chapter 9).** Any web system must respect legal norms and users and operators of such systems must be aware of the risks and the respective rights. Since, especially for young technology, it is not always easily predictable what possible conflicts with the law can arise from the use of a feature, previous research has discussed opportunities and risks of various of these new possibilities. Regarding scholarly publication tools, studies have dealt with the basic features of social tagging systems and with recommendations (see above). Thus, it remains to conduct a similar investigation on the possibility of reviewing in scholarly social bookmarking systems. Additionally to aspects of data protection, here, intricacy arises from users evaluating the work of others. Allowing



users to publicly criticize scholarly work – thus (at least implicitly) also its authors – bears the danger of misuse. Therefore, in Chapter 9, we first discuss the opportunities and risks of online rating systems in general and then focus on the special case of rating scholarly literature online – a process known as social peer review. Next to the discussion of various means to design such a feature in a socially compatible manner, we also discuss how it can be integrated into a social bookmarking system, where (in contrast to other systems, like digital libraries) the resources are contributed by users.

### 1.3 Summary of the Contributions of this Thesis

In this section, we summarize the main contributions of this thesis – to the analysis of research, to the investigation of social bookmarking systems and to research on recommendations within such systems.

**Analysis of Research Fields and Research Communities.** We analyze several research communities, using data from different stages of the publication life cycle. We conduct well-known scientometric analyses, and we introduce methodology from other research areas, such as formal concept analysis (Chapter 3), social network analysis (Chapters 3 and 4), and web log mining (Chapters 5 and 6) to the field of analyzing research and researchers. We introduce the conference guidance system Conferator, that assists researchers during the second phase of the publication life cycle, that is, during the dissemination phase, when the authors visit conferences to present their work (Chapter 4).

**Social Bookmarking Systems.** Through use studies on BibSonomy, we investigate how social bookmarking systems are used. For the first time, by exploiting web log data, we are able to analyze what users actually do in a tagging system. We focus on social interactions, connections between usage and retrieval (Chapter 5), and on the usage intensity of publications in a tagging system and their (future) scholarly impact, measured in citations (Chapter 6). Moreover, we discuss how online rating systems can be integrated into a social tagging system, and what can or should be done to design them legally and socially compatible, protecting those who rate and those who (or whose products) are rated (Chapter 9).

**Folksonomic Recommender Systems.** We extend the frequently used recommender systems benchmarking scenario that relies on graph-cores by introducing set-cores – a generalization of the notion of cores from graphs to arbitrary sets. Moreover, we critically scrutinize the consequences of relying on this setup and list several pitfalls (Chapter 7). Furthermore, we analyze the use of *FolkRank* for the recommendation of scholarly publications, and we extend it to allow the inclusion of additional data, beyond the folksonomy structure, which is *FolkRank*'s usual input (Chapter 8).



## Chapter 2

### Foundations and Related Work

In this chapter, we recall fundamental methods and discuss related work from research areas that are the most relevant to the topics in this thesis. Further literature will be mentioned in individual sections of each chapter. Section 2.1 is a collection of basic methodology and several means of analysis which we use in the thesis. Sections 2.2 through 2.4 present a broad overview of the context of the two main themes in this thesis. They also introduce some aspects in greater detail since they will be discussed or used later on. In Section 2.2, we discuss scientometrics, the field to which the analyses along the publication life cycle (📄) belong. Section 2.3 is concerned with social bookmarking systems and thus the context for the second theme of the thesis (📌). Section 2.4 also belongs to that theme, discussing recommender systems, specifically those used in social bookmarking systems. It is mainly relevant for the chapters in Part III.

When in one of the following chapters, notions from this chapter are used, we always refer back to the respective section here. Thus, readers may skip this Chapter now and come back to individual sections when referred to later on.

#### 2.1 Basics and Methodology

In this section, we introduce various theoretical or technical concepts and methods that are used or built on in the following chapters of this thesis.

##### 2.1.1 Analyzing Empirical Distributions

In various analyses, we discuss properties of empirical distributions. For instance, we compare three conference series through their citation distributions in Chapter 3, we analyze face-to-face contact distributions in Chapter 4, we compare tag popularity in posts and requests of a tagging system in Chapter 5, and we analyze correlations between publication usage and citations in Chapter 6. For these tasks, we fit the actual, empirical distributions to possible theoretical candidates, and we compute correlations between distributions from different datasets to get an impression on their relation towards each other. Therefore, the fitting of distributions and the computation of correlations is recalled in the following.

### Fitting Distributions

We briefly recall the methodology from Clauset et al. [2009] for fitting power-law distributions, as well as for testing those fits against other candidate distributions. Clauset et al. [2009] discussed several statistical properties of distribution fits and showed that their method is superior to other modes of fitting. We apply their methods on various occasions in this thesis, using the implementations by Clauset et al. [2009] and Alstott et al. [2014]. The task is to determine parameters  $\alpha \in \mathbb{R}$  and  $x_{min} \in \mathbb{R}$  such that a set of observations of some quantity (e.g., citation frequencies of publications) is likely to have been drawn from a distribution

$$p(x) \propto x^{-\alpha}, \text{ with } x \geq x_{min}.$$

The following procedure has been proposed by Clauset et al. [2009]:

**Fitting.** For all possible  $x_{min}$  (the empirically found values), a fit is computed by determining the parameter  $\alpha$  as a maximum likelihood estimator. From all fits, that  $(x_{min}, \alpha)$  combination is selected which minimizes the Kolmogorov-Smirnov statistic – the maximum distance between the cumulative distribution function of the observed empirical data and the fitted distribution (for  $x \geq x_{min}$ ).

**Uncertainty of the Fit.** To compute an uncertainty measure for the selected parameters, samples are chosen uniformly at random from the empirical data. Fits are determined for the samples and their standard deviation is taken as the measure of uncertainty for the original fit. Clauset et al. [2009] suggest to use 1,000 repetitions.

**Testing the Hypothesis of a Good Fit.** A goodness-of-fit (gof) test generates a  $p$ -value for the hypothesis that the observed data is actually drawn from the distribution determined through the fit. For the test, a number of synthetic datasets is drawn from the fit distribution, and then the Kolmogorov-Smirnov statistic between it and the fit is computed. The  $p$ -value counts the share of those cases where this statistic is higher than the Kolmogorov-Smirnov statistic between the original empirical data and the fit (i.e., the share of samples from the fitted distribution where the latter is a worse fit than it is for the original data). Thus a high  $p$ -value is an argument supporting the plausibility of the hypothesis, whereas a low  $p$ -value would be evidence against the power-law assumption. Clauset et al. [2009] suggest to use 2,500 synthetic datasets to compute  $p$  and to use  $p = 0.1$  as the threshold for rejecting a power law.

**Comparison to Other Candidate Distributions.** Even if the above test does not reject a power law, other distributions might actually be better fits for the data. Therefore, it suggests itself to compare the goodness-of-fit statistic of the power law to that of other distributions. For this purpose, Vuong’s closeness test [Vuong, 1989] is used to determine if one of two alternatives (the power-law fit or one other fitted

candidate distribution) is significantly better than the other. A  $p$ -value determines the significance of the difference in the goodness-of-fit. If it is small then the hypothesis that both candidates are equally good fits can be rejected.

### Correlation

Correlation coefficients are measures of dependence between two given series of measurements. Pearson's and Spearman's correlation coefficients are the standard approach to determine to what extent two random variables are dependent, given their observed distributions.

**Pearson's Correlation Coefficient.** Given two discrete random variables  $X$  and  $Y$  with sample pairs  $(x_1, y_1), \dots, (x_n, y_n)$  and their average values  $\bar{x}$  and  $\bar{y}$ , Pearson's correlation coefficient  $r$  is computed as

$$r(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

The range of  $r$  is the closed interval  $[-1; 1]$ , where the highest value  $r = 1$  denotes a perfect linear correlation between  $X$  and  $Y$ , and the lowest value  $r = -1$  denotes a perfect inverse linear correlation. A correlation value of  $r = 0$  indicates that there is no correlation between  $X$  and  $Y$ .

**Spearman's Rank Correlation Coefficient.** Given  $X$  and  $Y$  as above, Spearman's  $\rho$  works similar as Pearson's  $r$ , only before the coefficient is computed,  $x_i$  and  $y_i$  are replaced by their ranks  $\text{rank}(x_i)$  and  $\text{rank}(y_i)$  which they assume in the respective ordered lists of all  $X$  or all  $Y$  samples:

$$\begin{aligned} \rho(X, Y) &= r(\text{rank}[X], \text{rank}[Y]) \\ &= \frac{\sum_{i=1}^n (\text{rank}(x_i) - \overline{\text{rank}[X]})(\text{rank}(y_i) - \overline{\text{rank}[Y]})}{\sqrt{\sum_{i=1}^n (\text{rank}(x_i) - \overline{\text{rank}[X]})^2} \sqrt{\sum_{i=1}^n (\text{rank}(y_i) - \overline{\text{rank}[Y]})^2}}, \end{aligned}$$

with  $\text{rank}[X]$  and  $\text{rank}[Y]$  being the average ranks of  $X$  and  $Y$ , respectively. Other than Pearson's  $r$ , Spearman's correlation value determines how well the relationship of  $X$  and  $Y$  can be described using some monotone function (instead of testing for a linear relationship).

**Significance Testing for Correlations.** A  $p$ -value can be computed to test against the null hypothesis that  $X$  and  $Y$  are in fact not correlated. Given the correlation coefficient  $r$ , it states how likely it is to yield a result at least as extreme as  $r$ , assuming that the correlation coefficient for uncorrelated sequences follows a  $t$ -distribution. For details see [Sachs, 1988]. Thus, a low  $p$ -value would be a plausibility argument for actual correlation, while a high  $p$ -value suggests to presume that  $X$  and  $Y$  are uncorrelated.

### 2.1.2 Graphs

Graphs are the mathematical representation of networks, modeling one or more sets of entities together with connections between individuals from these sets. In this thesis, graphs will be used, for example, to model co-authorship relations and citation networks between authors or between authors and publications in Chapter 3, face-to-face contact networks in Chapter 4, and the user-resource-tag relation in social bookmarking systems, which we will use in Chapters 5 through 8. In these chapters, we analyze various properties of these graphs as well as particular subgraphs (cores in Chapter 7), and therefore, we recall the relevant notions in the following.

**Definition 2.1** (Graph). *A directed graph  $G = (V, E)$  is an ordered pair, consisting of a finite set  $V$  of vertices or nodes, and a set  $E \subseteq V \times V$  of edges. An undirected graph is defined accordingly, only here  $E$  denotes a set of two-element subsets of  $V$ . In a weighted Graph each edge  $e \in E$  is assigned an edge weight  $w(e)$  by some weighting function  $w: E \rightarrow \mathbb{R} : e \mapsto w(e)$ .*

A useful encoding of a graph's edge set  $E$  is the *adjacency matrix*  $A \in \mathbb{R}^{n \times n}$  where  $n = |V|$  is the number of nodes. Enumerating the nodes of  $V$  by  $1, 2, \dots, n$ ,  $A$  is defined through  $a_{ij} := 1$  if  $\{i, j\} \in E$  (or in a directed graph  $(i, j) \in E$ ) and  $a_{ij} := 0$  otherwise. For weighted graphs, the matrix can also include the edge weights, for instance, in an undirected graph  $a_{ij} := w(\{i, j\})$ . Two nodes  $u, v \in V$  are said to be *connected* if there is a path between them, that is, a series of nodes  $u = v_0, v_1, \dots, v_l = v$  with  $l \geq 1$  and  $(v_i, v_{i+1}) \in E$ , or respectively  $\{v_i, v_{i+1}\} \in E$  if  $G$  is undirected, for  $i = 0, \dots, l-1$ . The number of edges  $l$  denotes the length of the path. A *shortest path* between nodes  $u$  and  $v$  is such a path with minimal length. Several statistics are often used to describe a graph's properties, among those are:

**Density** is the ratio of the number of edges to the number of possible edges (i.e., the number of edges in a fully connected graph with the same number of nodes).

**The number of connected components** counts the vertex subsets in the partitioning of  $V$  that is induced through the connectedness of nodes.

**Average Path Length (APL)** is the mean length of shortest paths between any two nodes in the graph.

**The clustering coefficient** is the ratio of three cliques (three nodes where each is connected to both others) to the number of sets with exactly three nodes that form a connected subgraph.

Similarly to choosing subsets from a set, one can select subgraphs in a graph:

**Definition 2.2** ((Induced) Subgraph). *A subgraph of a graph  $G = (V, E)$  is a graph  $(W, F)$ , such that  $W \subseteq V$  is a subset of the nodes of  $G$  and  $F \subseteq E$  is a subset of the edges of  $G$ . Given a subset of nodes  $W \subseteq V$ , the maximal subgraph of  $G$  having  $W$  as its node set is said to be induced by  $W$  and denoted by  $(W, E|_W)$ .*

Note that the requirement of  $(W, F)$  being a graph means that the edges in the subgraph must connect only nodes from the subgraph's node set  $W$ , meaning

- $\forall e \in F : e \subseteq W$  if  $G$  is undirected, or respectively
- $\forall e \in F : e \in W \times W$  if  $G$  is directed.

The induced subgraph is the restriction of  $G$  to a subset of nodes where the remaining nodes are connected to each other in the same way as they are in  $G$ .

The above definitions describe networks in which each edge connects two entities. Using graphs for modeling other data structures, like folksonomies (see Section 2.3.1), can require edges that connect more than two entities. Such structures are called *hypergraphs*:

**Definition 2.3** ( $(k$ -dimensional) Hypergraph). *An undirected hypergraph  $H = (V, E)$  consists of a set of nodes  $V$  and a set of (hyper)edges  $E$  where  $E \subseteq \mathfrak{P}(V)$  is some subset of the power set of  $V$ . If for all  $e \in E$  holds  $|e| = k$ , for some  $k \in \mathbb{N}$ , then  $H$  is called a  $k$ -dimensional hypergraph.*

In this terminology, (regular) undirected graphs are two-dimensional hypergraphs. When a graph is used to model connections between different kinds of entities (e.g., between users and items or, like in the case of tagging systems, between users, tags, and resources) the node set is partitioned into subsets, one for each type of entity. When in such a graph with  $k$  entity types, each edge always connects one entity per type, the graph is called  $k$ -partite:

**Definition 2.4** ( $k$ -partite Hypergraph). *A  $k$ -dimensional hypergraph  $H = (V, E)$  is called  $k$ -partite, if its node set can be partitioned into  $k$  non-empty sets  $V_1, V_2, \dots, V_k$ , such that for each hyperedge  $e \in E$  holds  $|V_i \cap e| = 1$  for all  $1 \leq i \leq k$ .*

In such graphs, we write for convenience  $e = (v_1, v_2, \dots, v_k)$ , with  $v_i \in V_i$  ( $1 \leq i \leq k$ ), to denote  $e = \{v_1, v_2, \dots, v_k\}$ .

While graphs are as such a relatively simple data structure, a large number of real-world problems can be projected onto graphs, to be tackled there efficiently using graph mining methodology. In the following, we recall a few typical problems that can be approached in graphs:

### Important Nodes

One typical task in network analysis is the identification of exceptionally *interesting*, *central*, or *important* nodes. For example, in a social network, like between co-authors (Chapter 3) or between participants of a conference (Chapter 4), it is often interesting to find the key players in the respective community. Various measures of importance have been proposed and usually, the use case determines which of them is the most appropriate. A set of various centrality measures is collected in [Koschützki et al., 2005]. According to their Definition 3.2.1, a *centrality index*  $c$  must be a real-valued

function on the node set of a graph that respects graph isomorphisms. In the following, we recall those measures that are of relevance in this thesis. When we refer to a graph, we assume it to be denoted like in Definition 2.1.

**Node Degree and Node Strength.** The most simple notion of the importance of a node  $v \in V$  is its *degree*  $\deg(v)$ , which counts the number of edges that contain  $v$ . In directed graphs, one can additionally define the *outdegree*  $\deg^{\text{out}}(v)$  as the number of edges that start at  $v$ , formally  $\deg^{\text{out}}(v) := |\{(v, u) \in E \mid u \in V\}|$ , and the *indegree*  $\deg^{\text{in}}(v)$  as the number of edges that point to  $v$ :  $\deg^{\text{in}}(v) := |\{(u, v) \in E \mid u \in V\}|$ . In weighted graphs, one can define *strength*  $\text{str}(v)$ , *instrength*  $\text{str}^{\text{in}}(v)$ , and *outstrength*  $\text{str}^{\text{out}}(v)$  of a node  $v$  analogously to degree, indegree, and outdegree, by adding up the respective edge weights instead of just counting edges. For instance, the instrength is computed as

$$\text{str}^{\text{in}}(v) := \sum_{(u,v) \in E} w((u, v)).$$

**Betweenness.** The betweenness centrality  $\text{bet}(v)$  of a node  $v \in V$  measures the importance of a node by looking beyond its local properties (e.g., counting its neighbors like the degree). Betweenness considers shortest paths between two points in the graph. A central node is one that lies on many shortest paths between many pairs of nodes. The measure has been proposed by Freeman [1977] and is formally defined as:

$$\text{bet}(v) := \sum_{u \neq w \in V \setminus \{v\}} b_{u,w}(v) \quad \text{with} \quad b_{u,w}(v) := \begin{cases} \frac{g_{u,w}(v)}{g_{u,w}} & \text{if } u \text{ and } w \text{ are connected} \\ 0 & \text{otherwise,} \end{cases}$$

where  $g_{u,w}$  is the number of all shortest paths (geodesics) between the nodes  $u$  and  $w$ , and  $g_{u,w}(v)$  is the number of such paths that contain  $v$ .

**PageRank.** *PageRank* [Brin and Page, 1998] is an eigenvector-based measure, originally developed to measure the importance of web pages based on the link structure of the World Wide Web. The main idea of the ranking is that important nodes are pointed to by other important nodes. To assign a score to each node in a graph, a linear equation system is (iteratively) solved which combines the graph structure (using the graph's adjacency matrix) and a probabilistic component. The equation models the behavior of a "random surfer" who either follows links on a web page to get to another or who decides to jump to some other website with the probability  $(1 - d)$ ,  $0 < d < 1$ . With web pages as nodes and links as directed edges, this model is easily transferred to arbitrary directed graphs, yielding the following iteration:

$$\vec{c} \leftarrow dA^T \vec{c} + (1 - d)\vec{\mathbf{1}},$$

$\vec{\mathbf{1}}$  being a vector with  $|V|$  entries, all equal to one,  $\vec{c}$  being a vector that holds for every node its centrality score, and  $A$  being the adjacency matrix with its rows normalized,



meaning, for  $u, v \in V$  :  $a_{uv} := \frac{1}{\text{deg}^{\text{out}}(u)}$  if  $(u, v) \in E$  and  $a_{uv} := 0$  otherwise. *PageRank* is also the basis for the recommendation algorithm *FolkRank* [Hotho et al., 2006c], which we will use and discuss in Part III of this thesis.

**HITS.** Based on a similar idea as *PageRank*, Kleinberg [1999] proposed an algorithm that discovers *hubs* and *authorities* in a directed graph. A hub is a node that points to many good authorities and an authority is a node that is pointed to by many good hubs. For each node  $v$ , the *HITS* algorithm computes iteratively alternating its authority score from the hub scores of those nodes pointing to  $v$  and its hub score from those nodes it points to. Like the *PageRank* computation, the task can be formulated as an eigenvector computation problem.

## Communities

The task of identifying communities in a graph has a wide range of applications. In his survey on the topic, Fortunato [2010] presents examples for the detection of communities among others in computer science, in protein-protein interaction networks, and in person-interaction networks. To the latter belong collaboration networks among researchers, which we will discuss in Part I, where collaboration is expressed through co-authorship or through conversations. Fortunato [2010] also notes, that the problem of community discovery is not clearly defined and in fact that the definition of a community is often given as the product of a community detection algorithm.

Intuitively – though vague –, communities are certain subsets of some larger set of entities, such that the members of a subset are somewhat more related or more similar to each other than they are to others. In undirected graphs, this intuition can be expressed with edges: One expects a community to be a set of nodes which are more densely connected with each other than with the other nodes in the graph. In this thesis, we understand the task of community discovery as the task of generating a graph clustering, which is a partition<sup>1</sup> of the graph’s node set into subsets. Therefore, we can use the words “cluster” and “community” interchangeably. For a given set of communities  $C$ , a *community allocation* is a mapping  $c : V \rightarrow C : v \mapsto c(v)$ , that assigns to each node  $v$  the community  $c(v)$  it belongs to.

Given a community allocation, naturally the question arises, how well these communities are aligned with the graph structure – following the above idea that communities should be well-connected internally and rather sparsely connected to nodes from other communities. Scripps et al. [2007b] expressed this intuition with a pair of alignment metrics  $p$  and  $q$ .

<sup>1</sup>Elsewhere, for example in [Atzmueller et al., 2016a], communities can overlap and thus entities can be part of more than one community at the same time, or must not belong to any community at all. Such models are useful, when more than one criterion is available for forming communities, like in a social network, where users can be interested in various topics and thus make contact to other users who might share the interest for one or several, but not necessarily for all topics. The chosen model of communities depends on the use case.

**Definition 2.5** (Community Alignment, cf. [Scripps et al., 2007b]). *Given an undirected graph  $G = (V, E)$ , a set of communities  $C$ , and a community allocation  $c : V \rightarrow C : v \mapsto c(v)$ , the alignment of the allocation is measured with the statistics  $p$  and  $q$ , where*

$$p := \frac{|\{\{u, v\} \in E \mid c(u) = c(v)\}|}{|E|} \quad \text{and} \quad q := \frac{|\{\{u, v\} \subseteq V \mid \{u, v\} \notin E, c(u) \neq c(v)\}|}{|\{\{u, v\} \subseteq V \mid \{u, v\} \notin E\}|}.$$

The alignment determines how well the edges in a graph fit the given community allocation:  $p$  is the share of edges that connect nodes within a community, and  $q$  is the share of unconnected node pairs, where both nodes belong to different communities. A high result for  $p$  means that connected nodes are very likely to belong to the same community and a high result for  $q$  means that two unconnected nodes are likely to belong to different communities. Ideally, both values are equal to one, in which case the communities are isolated cliques.

Since the above community alignment assessment consists of two metrics, it is not suitable for algorithms that optimize a community allocation towards one fix target function. For such purposes, a single measure must be applied. The measure *modularity*, proposed by Newman and Girvan [2004] and extended to weighted graphs in [Newman, 2004], is the most popular measure among them, and various community detection algorithms use modularity as the target to be optimized. Modularity computes the difference between the fraction of a graph's edges that connect nodes within the same community and the expected value for that fraction in a random graph with the same number of nodes and the same degree distribution:

**Definition 2.6** (Modularity, cf. [Newman and Girvan, 2004, Fortunato, 2010, Newman, 2004]). *Given an undirected graph  $G = (V, E)$  with adjacency matrix  $A$ , a set of communities  $C$ , and a community allocation  $c : V \rightarrow C : v \mapsto c(v)$ , the modularity  $\text{MOD}(c)$  of the allocation is the value*

$$\text{MOD}(c) := \frac{1}{2|E|} \sum_{u \neq v \in V} \left( A_{u,v} - \frac{\deg(u)\deg(v)}{2|E|} \right) \delta(c(u), c(v)),$$

with  $\delta(c(u), c(v)) = 1$  if  $c(u) = c(v)$  and  $\delta(c(u), c(v)) = 0$  otherwise. In a weighted graph with the weighting function  $w : E \mapsto \mathbb{R}$ , the weighted modularity is defined as above with  $A_{u,v} := w(\{u, v\})$  and the degree replaced by strength.

## Node Roles

A task that is related to the discovery of important nodes is that of determining roles of nodes: Each node is labeled with a particular role that describes its position or some of its structural features in the graph. We will study node roles in networks of co-authorship (Chapter 3) and of face-to-face contacts (Chapter 4). Particularly, in Chapter 4, we will group a conference's participants by their academic position and

study which roles participants in different positions assume. To determine community-based roles, we use methodology by Scripps et al. [2007a,b]. A special feature of their approach is that it allows the computation of roles not only when a community allocation is already given, but it also provides a solution for the case where no communities are determined. Scripps et al. [2007a,b] created a set of four roles that nodes can play in a network: *Loner*, *Bridge*, *Big Fish*, and *Ambassador*. These roles are assigned based on two measures, the node degree and a community metric. The latter is a measure that assesses how many communities a node connects. For a given allocation of nodes to communities, this measure can simply be counted [Scripps et al., 2007a]. If, however, no particular community allocation is available, Scripps et al. [2007b] provide an estimate called `rawComm` for that quantity. For a node  $u \in V$  it depends on the neighborhood  $N(u)$  of  $u$  (the set of nodes adjacent to  $u$ ) and on the alignment metrics  $p$  and  $q$  from Definition 2.5. It is computed as

$$\text{rawComm}(u) := \sum_{v \in N(u)} \tau_u(v),$$

where  $\tau_u(v)$  is the contribution of a neighbor  $v$  of  $u$ :

$$\tau_u(v) := \frac{1}{1 + \sum_{v' \in N(u)} (p \cdot \delta_{\{v,v'\} \in E} + (1-q) \cdot (1 - \delta_{\{v,v'\} \in E}))},$$

with  $\delta_{\{v,v'\} \in E}$  being the Kronecker symbol (i.e.,  $\delta_{\{v,v'\} \in E} = 1$  if there is an edge between  $v$  and  $v'$  and  $\delta_{\{v,v'\} \in E} = 0$  else). The score  $\text{rawComm}(u)$  is the expected value for the number of communities  $u$  connects in the graph, given  $p$  and  $q$ . If no community allocation is given,  $p$  and  $q$  have to be estimated. Scripps et al. [2007a] suggest setting  $p = q = 1$ .

For a node  $u$ , given its normalized degree  $\text{ndeg}(u)$  and its `rawComm` score, one of four roles is assigned to  $u$  following this rule:

$$\text{role}(u) := \begin{cases} \text{Ambassador} & \text{if } \text{ndeg}(u) \geq s, \text{rawComm}(u) \geq t \\ \text{Big Fish} & \text{if } \text{ndeg}(u) \geq s, \text{rawComm}(u) < t \\ \text{Bridge} & \text{if } \text{ndeg}(u) < s, \text{rawComm}(u) \geq t \\ \text{Loner} & \text{if } \text{ndeg}(u) < s, \text{rawComm}(u) < t, \end{cases} \quad (2.1)$$

with  $s$  and  $t$  being thresholds that have to be set. Scripps et al. [2007b] suggest to use  $s = t = 0.5$  after normalizing `rawComm` to the interval  $[0, 1]$ . *Ambassadors* are characterized by high scores in both degree and `rawComm`, which is interpreted as being important (high degree) and at the same time connecting many communities in the graph (high `rawComm`). A *Big Fish* has connections to a lot of other nodes and is thus important, however, mostly within only few communities. *Bridges* connect communities, however, not as many as ambassadors. Finally, *Loners* are those with low scores in both measures.

## 2.2 Analysis of Research Fields and Research Communities

In this section, we discuss the larger context to which the contributions of the first theme of this thesis (I) belong, that is, the field of informetrics, or more specifically, scientometrics. Moreover, we recall previous findings on citation distributions and impact analyses, as we will apply such methodology in our studies, mainly in Parts I and II. For the analysis of research, researchers, scholarly publications, and scholarly processes, mostly statistical methods are employed to conduct quantitative analyses. They allow comparisons of different forms of impact of individual publications or even of researchers. They can be relevant in the allocation of funding, in the appointment of professorships, or simply to spark competition among researchers. Apart from the evaluation perspective, analyzing research can also support researchers in their daily work, which is the goal of this thesis. For example, such analyses can help by identifying the most relevant publications and players in an academic community, current (or even future) hot topics, or related work within the huge corpus of available literature.

Quantitative studies of research (often of scholarly literature) can be found under the keywords *scientometrics* or *bibliometrics*, *webometrics*, *informetrics*, *cybermetrics*, and occasionally others (see [Björneborn and Ingwersen, 2004] for further examples). It is not easy to distinguish between these terms as they have been proposed in different contexts – historical roots of the terms biblio-, sciento-, and infor-metrics are visited by Brookes [1990] –, and they have often been used synonymously. To categorize the analyses presented in this thesis (Parts I and II), we use the framework of notions introduced by Tague-Sutcliffe [1992] and Björneborn and Ingwersen [2004]. Tague-Sutcliffe [1992] gave the following three definitions:

**Bibliometrics** is “the study of the quantitative aspects of the production, dissemination and use of recorded information.”

**Scientometrics** is “the study of the quantitative aspects of science as a discipline or economic activity.”

**Informetrics** is “the study of the quantitative aspects of information in any form, not just records or bibliographies, and in any social group, not just scientists.”

Björneborn and Ingwersen [2004] added to these notions by proposing two further definitions:

**Webometrics** is “the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches.”

**Cybermetrics** is “the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the whole Internet drawing on bibliometric and informetric approaches.”

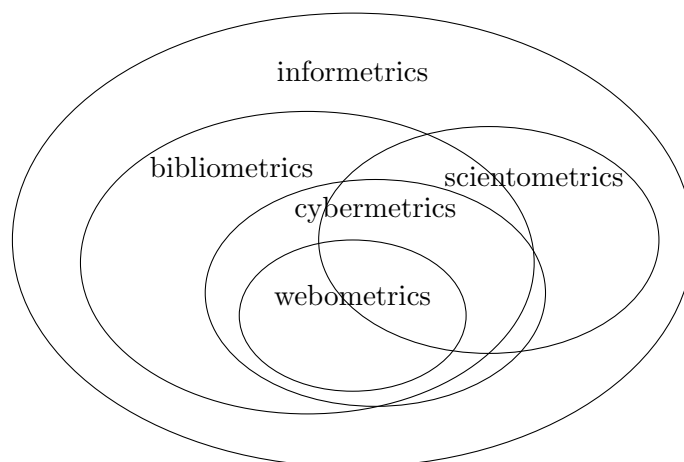


Figure 2.1: Relationships between informetrics and four of its subdisciplines – scientometrics, cybermetrics, webometrics, and bibliometrics – according to Björneborn and Ingwersen [2004]. The diagram can be found similarly there.

Furthermore, Björneborn and Ingwersen [2004] discussed and visualized the relationships between the five different fields in the diagram that is shown here in Figure 2.1. Following the definitions above it shows informetrics as the broad all-encompassing field with the other four as sub-fields. Webometrics is a subset of both cybermetrics and bibliometrics. Bibliometrics and scientometrics overlap when scholarly information is analyzed. In contrast to scientometrics, bibliometrics also covers non-scholarly information, whereas scientometrics, unlike bibliometrics, is not restricted to recorded information, but can deal with all kinds of research activities. For more details on the various intersections or differences between these fields see [Björneborn and Ingwersen, 2004]. More recently, the term *altmetrics* [Priem et al., 2010] has become popular. It describes alternative metrics (as opposed to metrics relying on citations) for the impact of citations. Altmetrics are quickly available metrics based on data from the usage phase of the publication life cycle, utilizing links, downloads, mentions in social media, bookmark posts, or discussions related to a publication. They are a relatively young subfield of bibliometrics and scientometrics, and they are related to webometrics [Priem et al., 2010].

Following the notions above, our work in Parts I and II of this thesis can be classified as part of scientometrics, as it deals solely with scholarly data. Some of the work belongs to bibliometrics where we analyze recorded information like metadata of publications or citations between them (Chapters 3, 5, and 6). In Chapter 6, we approach the field of altmetrics. The studies in Part III of this thesis belong to the fields of social bookmarking and recommender systems, which we will review in Sections 2.3 and 2.4, respectively.

### 2.2.1 Analyzing Citations

A central role in the evaluation of publication impact is played by the citations a publication has received. Thus, citations have been the subject of various studies, including investigations of their distribution, of their use in impact measures, or of the influence of self-citations. We will compare citation distributions in Chapters 3 and 6 and therefore, here, we recall previous findings concerning these topics.

#### Citation Distributions

It is often presumed that citation counts follow power-law distributions. Clauset et al. [2009] computed fits for citations in a large dataset, comprising articles that are listed in the Science Citation Index and that have been published in 1981, and their citations until June 1997. The exponent  $\alpha$  of the best fit to a power-law distribution (see Section 2.1.1) is  $\alpha = 3.16$ , and the fit starts at  $x_{min} = 160$  – thus only articles with at least that many citations are covered. The fit's  $p$ -value  $p = 0.2$  does not suggest to reject the assumption of a power law. They also find that fits to an exponential or stretched exponential are a worse fits, while a power law with exponential cut-off or a lognormal distribution might be more suitable, however, the difference is not significant.

Albarrán and Ruiz-Castillo [2011] conducted a study, using the same methodology, on a sample of articles in the Web of Science, covering five years and a five-year citation window. Particularly for mathematics and computer science – the two disciplines which the articles studied in this thesis (citation distributions play a role in Chapters 3, 5, and 6) most likely belong to –, they observed the  $p$ -values  $p = 0.614$  and  $p = 0.672$ , respectively. Thus, the assumption of a power-law distribution cannot be rejected. With  $\alpha = 3.83$ ,  $x_{min} = 20$  and  $\alpha = 2.92$ ,  $x_{min} = 18$ , the fits cover only about 0.86% and 2.22% of the articles, respectively. Over all investigated disciplines the results were varying, including a few fields where a power-law distribution was rejected. The exponent  $\alpha$  assumed values between  $\alpha = 2.92$  and  $\alpha = 5.05$ , and the threshold  $x_{min}$  varied between  $x_{min} = 18$  and  $x_{min} = 152$ .

Brzezinski [2015] used all articles from 27 Scopus major subject areas and a citation window of five years, and fitted citation counts to power laws according to [Clauset et al., 2009]. In their study, for mathematics and computer science, the power-law assumption had to be rejected ( $p$ -values of  $p = 0.012$  and  $p = 0.000$ ). The fits covered 3.0% and 2.1% of the papers with parameters  $x_{min} = 24$  and  $x_{min} = 26$  and  $\alpha = 3.11$  and  $\alpha = 2.78$ . Similar to Albarrán and Ruiz-Castillo [2011], they noted many different values for different disciplines (e.g.,  $\alpha$  between  $\alpha = 2.78$  and  $\alpha = 4.69$ ). For all subject areas, power laws yield better fits than exponential distributions, yet for many areas, Weibull or lognormal distributions and power laws with exponential cut-off were better fits than the plain power-law distribution.

Overall, the results are somewhat inconclusive. It is obvious that the distributions (quality of the fits and exponents) depend heavily on parameters like the selected

research area and the sample of considered publications. In cases where the statistical tests do not suggest to reject the fit, it often covers only a small part of the observed data (the publications with particularly many citations).

### Citation-based Impact Measures

One of the main goals in citation analysis is the measurement of scholarly impact. The basic assumption is that citing a publication is a form of acknowledging its influence. Since we will use the journal impact factor to compare the success of conferences in Chapter 3, and since we compare citation-based impact and altmetrics in Chapter 6, here, we repeat the necessary basics and recall arguments for and against the inclusion of self-citations in such studies.

Citation-based impact measures can be distinguished by their domain: (i) *Article-level metrics* measure the impact of individual publications, for instance, by counting all citations the publication received. (ii) *Author-level metrics* assess the overall impact of an author’s work (e.g., the h-index [Hirsch, 2005]), while (iii) *Journal-level metrics* (e.g., the journal impact factor, see below) compare the success of journals. Journal-level metrics can simply be generalized to measure the impact of other venues like conferences (as in Section 3.5.1 for individual editions of FCA-minded conferences). Similarly, author-level metrics can be used as journal impact measures by replacing the authorship relation by the “published-in” relation.

**Journal Impact Factor.** The journal-level metric that is most well-known is the journal impact factor, proposed by Garfield [1972]. It is computed annually, using Web of Science data, and it is often reported by the journals themselves as a sign for their reputation. For a journal, its impact factor in year  $n$  is computed as the number of citations in year  $n$  to articles from the years  $n - 1$  and  $n - 2$ , divided by the number of articles published in these two years. We will use this metric in Chapter 3 to compute and compare the impact of three conference series.

**Self-Citations.** Self-citations (authors citing their own work) are a controversial subject in the scientometrics literature. It has been debated whether or not to include them in citation analyses. Since in two of our studies (Chapters 3 and 6) we conduct such analyses, in the following we recall some of the arguments of that debate. While citing oneself is natural [Phelan, 1999, Tagliacozzo, 1977] when continuing one’s own work, there might be other motives, like improving one’s own (citation-based) scores. Bonzi and Snyder [1991] asked 51 authors about their motivation to use particular citations: The most significant differences between self-citations and others were that authors used the former more often to refer to “earlier work on which current work builds” and to “establish writer’s authority in the field”. None of the authors checked the reason “raise citation count” for any citation. Aksnes [2003] found that the ratio of self-citations among received citations is particularly high for rarely cited articles. They attribute this tendency to the facts that (i) it is practically difficult to cite own work

more than only a few times and (ii) for a paper with only few citations it is most likely the authors that care for its content. However, Aksnes [2003] also notes that among publications that received only one citation, the majority of these are not self-citations. Overall, Aksnes [2003] concludes that for studying individual contributions one should consider possible effects of self-citations. Similarly argues Phelan [1999], who finds that among particularly highly cited authors, the correlation between the total number of citations and the number of non-self-citations is very strong ( $r = 0.925$ ), but for comparing the impact of individual authors “removing self-citations is an important prerequisite”. Finally, Thijs and Glänzel [2006] suggest to present citation-based impact measures both including and excluding self-citations.

In practice, self-citations are often included in impact measures. For instance, the annual Thomson Reuters Journal Citation Reports compute their metrics with all citations;<sup>2</sup> Google Scholar reports its author-level metrics, like overall citation count, h-index, and i10-index, including self-citations as well. The reasons for that may be manifold and we can only speculate about them: (i) It is not entirely clear that self-citations should not be considered as indicators for scientific impact. Aksnes [2003] mentions the case when a publication with more than one author is being cited in another paper, that shares only some of the authors of the first paper. From the perspective of an author of only the first paper, this would not be a self-citation. (ii) Excluding self-citations is technically non-trivial: Reference sections of an article are difficult to parse and sometimes, data is missing or erroneous. Our own experiences with creating a corpus (in Chapter 3) are that a lot of (manual) work must be spent on cleaning and completing such data to properly identify and filter self-citations. (iii) Removing self-citations makes it harder to compare metrics, since the mode of excluding self-citations might be different in other datasets due to the previous two reasons. In this thesis, we work with citations in Chapters 3 and 6, and we discuss the handling of self-citations there (Sections 3.4.2, 3.5, and 6.4.2).

By and large, citation-based impact measures can help estimate the scientific impact a publication, an author, or a venue has. Many such metrics (and their variants excluding or including self-citations) can be computed, provided the availability of the citation data. With the rise of altmetrics, even more measures can be constructed, using data from the social web instead of citations. However, these methods are not without pitfalls. Phelan [1999] collected several critical aspects – ranging from different citation practices over a language bias to technical difficulties – and therefore concluded: “Bibliometric analysis is but one tool among many, and so it should remain.” In the same spirit, the Leiden Manifesto [Hicks et al., 2015] lists as its first principle that “quantitative evaluation should support qualitative, expert assessment.”

---

<sup>2</sup>Only when a journal contains large shares of citations to its own articles, it is banned from the reports. For details see [http://admin-apps.webofknowledge.com/JCR/static\\_html/notices/notices.htm](http://admin-apps.webofknowledge.com/JCR/static_html/notices/notices.htm)



## 2.3 Social Bookmarking Systems

Social bookmarking systems are the central topic in this thesis, particularly in Chapters 5 through 9. Therefore, in this section, we introduce the fundamental idea of these systems, the underlying data structure – called *folksonomy* –, and the social bookmarking system BibSonomy, which will serve as use case in most of our studies. We also point to previous directions of research on such systems, indicating the broad interest they have sparked in our community.

The idea of social bookmarking is that people (i.e., the users of social bookmarking systems) collect and annotate resources online. Resources are annotated with freely chosen keywords, called tags. Usually, these annotations are publicly visible and thus a corpus of annotated resources emerges from the users' collaborative efforts. Therefore, these systems also go by the name of *collaborative tagging systems*. In such systems, tags can be used to retrieve the stored resources, by browsing through the tags (usually displayed in tag clouds) or by using tags in search queries. Tagging thus provides a benefit both for the users who assign tags to their resources, serving as a light-weight knowledge management system, as well as for all others, who can browse other users' content.

The term *social bookmarking* reflects (i) the idea of social sharing, when users publicly tag resources, and thus make them available to or at least retrievable for others, and (ii) the idea of bookmarking, the process of collecting resources for later retrieval. The term bookmarking is often used particularly for web pages. However, bookmarking systems can allow collecting any type of resource, like publication references, photos, music, or videos; basically, “anything with a URL” [Vander Wal, 2007].

Social bookmarking systems have been created for various resource types. Probably the most well-known system is Delicious,<sup>3</sup> where users share bookmarks to web pages. Delicious started in 2003 (as del.icio.us) and by the time of writing contains more than one billion bookmarked links.<sup>4</sup> The system Flickr<sup>5</sup> allows tagging images, on YouTube, users tag their videos (at the time of writing, YouTube has made these tags invisible in the web interface, yet it is acknowledged that they are used to help users find videos<sup>6</sup>), music can be tagged on last.fm, and on CiteULike, scientists can collect references to scientific articles. The social bookmark and publication sharing system BibSonomy allows bookmarking web pages and publication references. We make use of BibSonomy data in several chapters and therefore describe this system in greater detail below.

As dynamic web systems that contain almost only user-generated content, social bookmarking systems belong to the family of Web 2.0 applications (for details see [Peters, 2009, Chapter 1]). They are easy to use and fast: All a user needs to do, is to come up with some suitable (in his or her opinion) keywords and enter them. Even

---

<sup>3</sup><https://delicious.com/>

<sup>4</sup><http://delicious.com/about> (accessed June 1, 2015)

<sup>5</sup><http://www.flickr.com/>

<sup>6</sup><http://support.google.com/youtube/answer/146402> (accessed June 1, 2015)

that task is often assisted through tag recommendations, which will be a topic in this thesis and which is discussed in more detail in Section 2.4.1.

Meanwhile, new web systems have become popular, offering some advantages over tagging systems in some areas. For example, the microblogging service Twitter allows its users to quickly share links to web pages, simply by including them in a tweet. Thus, when long-term retrievability in the own collection is not required, users may prefer this way of sharing links over a dedicated bookmarking system. We assume that due to such developments, the number of active social bookmarking systems has shrunk. Previously popular social bookmarking systems, like Mister Wong (for web pages) or Connotea (for publication references), have been discontinued. Still, alone for web pages, the Open Directory Project lists 27 active bookmarking systems.<sup>7</sup>

On the other hand, tagging has found its place as a secondary feature in many systems: Twitter uses *hashtags* to generate topic-focused streams, Mendeley has integrated tagging into their product, a combination of social network and document manager, blogging software allows users to add tags to articles (e.g., Wordpress<sup>8</sup>), and web shops allow users to tag products (e.g., the computer game platform Steam<sup>9</sup>). Peters [2009] lists more applications in e-commerce, libraries, museums, or social networks. Finally, social bookmarking tools are applied in companies as light-weight knowledge management systems. A list of such systems can be found on Wikipedia.<sup>10</sup> Particularly, the system Dogear has also been subject to scientific analysis [Millen and Feinberg, 2006, Millen et al., 2007]. These systems are, however, usually only available within the Intranet of a company and not publicly available.

In the remainder of this section, we first recall a formal model for folksonomies – the data structures that emerge in social tagging systems and that are the basis for investigating social tagging. Then, we describe the social bookmarking system BibSonomy, which serves as primary use case in our investigations of the usage of such systems. Eventually, we present several examples of previous work on collaborative tagging systems, showing the relevance of these systems for various research directions.

### 2.3.1 Folksonomies

A notion that is sometimes used interchangeably with social bookmarking system is *folksonomy*, a term coined by Vander Wal [2007]. It combines the words *folk* and *taxonomy*, following the observation that the community of a tagging system’s users – the folk – create an informal and overlapping classification for a catalog of resources. Particularly, in contrast to taxonomies, a resource can be assigned to several “categories” (i.e., it can be annotated with several tags), and the categories (tags) are freely chosen by users rather than being selected from a pre-defined set by an expert.

---

<sup>7</sup>[http://www.dmoz.org/Computers/Internet/On\\_the\\_Web/Web\\_Applications/Bookmark\\_Managers/](http://www.dmoz.org/Computers/Internet/On_the_Web/Web_Applications/Bookmark_Managers/) (accessed June 1, 2015)

<sup>8</sup><http://en.support.wordpress.com/posts/tags/> (accessed June 1, 2015)

<sup>9</sup><http://store.steampowered.com/tag/> (accessed June 1, 2015)

<sup>10</sup>[http://en.wikipedia.org/wiki/Comparison\\_of\\_enterprise\\_bookmarking\\_platforms](http://en.wikipedia.org/wiki/Comparison_of_enterprise_bookmarking_platforms)

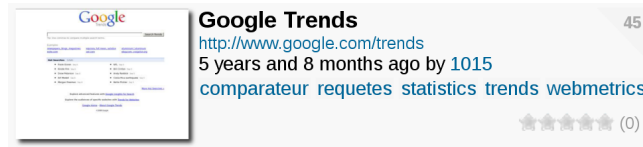


Figure 2.2: An example post from BibSonomy for the user with ID 1015 and the resource <http://www.google.com/trends>.

Rather than the tagging system itself, the term folksonomy describes the result of the tagging process and thus the data structure underlying these systems.

A folksonomy is created through the tagging activities of a tagging system's users. Every time users annotate resources, they create *tag assignments*. The collection of all tag assignments of a single user constitutes his or her *personomy*. More formally:

**Definition 2.7** (Folksonomy, cf. Definition 1 in [Hotho et al., 2006c]). *A folksonomy is a quadruple  $\mathbb{F} := (U, T, R, Y)$ , where  $U$ ,  $T$ , and  $R$  are finite sets, whose elements are called users, tags and resources, respectively, and  $Y \subseteq U \times T \times R$  is a ternary relation between them, whose elements are called tag assignments.*

*The personomy  $\mathbb{P}_u := (T_u, R_u, I_u)$  of a given user  $u \in U$  is the restriction of  $\mathbb{F}$  to  $u$  with  $I_u := \{(t, r) \in T \times R \mid (u, t, r) \in Y\}$ ,  $T_u := \pi_1[I_u]$ , and  $R_u := \pi_2[I_u]$ , where  $\pi_i$  denotes the projection on the  $i$ -th dimension.*

This definition is different to that in [Hotho et al., 2006c] as it omits a fifth component of a folksonomy, namely, a supertag/subtag relation. Since we do not use that in this work, we have excluded it here.

When users add a resource to their collection, they create a post in the system. Each post contains the user who created it (who owns it), the resource, and a set of tags. The post is thus a composite of several tag assignments. Viewing the data in a tagging system as a collection of posts is usually the way a folksonomy is perceived by users (cf. Figure 2.2).

**Definition 2.8** (Post). *A post in a folksonomy  $\mathbb{F} := (U, T, R, Y)$  is a triple  $(u, T_{ur}, r)$ , with  $u \in U$ ,  $r \in R$ , and  $T_{ur} \neq \emptyset$ , where  $T_{ur} := \{t \in T \mid (u, t, r) \in Y\}$ .*

A folksonomy  $\mathbb{F} := (U, T, R, Y)$  can be understood as a tripartite hypergraph  $G = (V, E)$  with users  $U$ , tags  $T$ , and resources  $R$  as the three partitions of the node set  $V = U \cup T \cup R$ . Each tag assignment  $(u, t, r) \in Y$  constitutes a (tri-)edge  $\{u, t, r\} \in E$  in the graph, connecting user  $u$ , tag  $t$ , and resource  $r$  with each other.

### 2.3.2 The Social Bookmarking System BibSonomy

BibSonomy is a social bookmarking system where users tag publication references and website bookmarks. It has not only attracted thousands of researchers who use it to manage their publication collections, but also the interest of the research community

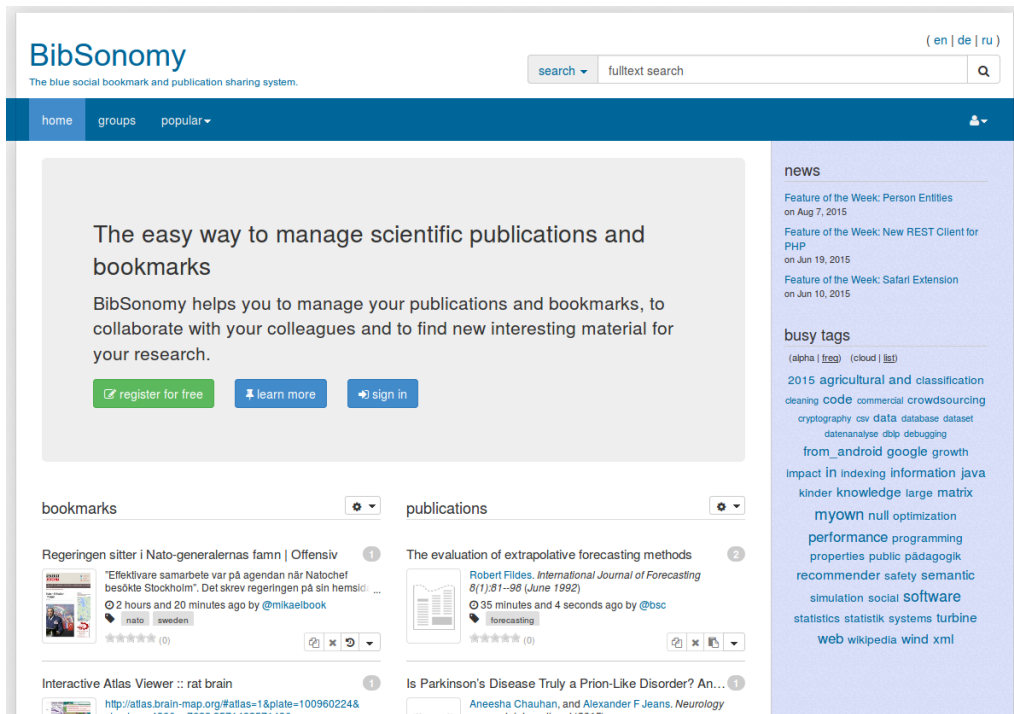


Figure 2.3: The landing page of the blue social bookmark and publication sharing system BibSonomy. Visible is the blue ribbon containing the navigation menu, BibSonomy’s welcome window, and the characteristic two-column view with bookmark posts (left) and publication posts (right). The blue sidebar contains news and the tag cloud.

as a subject to study. In this thesis, we will use BibSonomy data as our use case in various analyses. Therefore, in this section, we introduce BibSonomy in greater detail. A minimal familiarity with its basic features is relevant to follow the usage analyses in Part II.

The service BibSonomy started publicly in 2006 and has since been continuously developed further. The software is open source and available at Bitbucket.org.<sup>11</sup> A description of its architecture and various features has been provided by Benz et al. [2010a]. Figure 2.3 shows the BibSonomy landing page, always displaying the most recent posts that have been added. To introduce BibSonomy in greater detail, we use the taxonomy for characterizing tagging systems by their design, proposed by Marlow et al. [2006], and describe BibSonomy along its seven dimensions:

- *Tagging Rights*: BibSonomy allows users to tag any resources (in contrast to systems where users tag only content they created themselves). Users can edit and delete the tags they added but not those of others. Tagging a resource implies the creation of a post, that is added to the own collection (personomy).

<sup>11</sup><http://bibsonomy.bitbucket.org/>

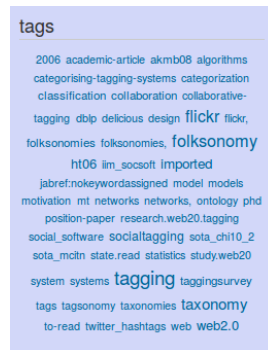


Figure 2.4: The tag cloud is shown in the sidebar of a publication’s details page in BibSonomy. It shows all tags that users have assigned to posts of that publication, in this example [Marlow et al., 2006].

BibSonomy provides an overview page for each resource that lists all the tags users have assigned to it, as well as the different posts containing the resource. For example, the publication by Marlow et al. [2006], that introduced this taxonomy, has been posted to BibSonomy by several users, resulting in a relatively large set of tags, visible in the sidebar of its overview page,<sup>12</sup> shown in Figure 2.4.

- *Tagging Support*: BibSonomy supports *suggested tagging* (the other alternatives in this dimension are *viewable*, where users are shown all tags that have previously been assigned to this resource, and *blind*, without assistance). Its multiplexing tag recommender is described in [Jäschke et al., 2009] and has been the basis for the online evaluation in the ECML PKDD Discovery Challenge 2009.<sup>13</sup>
- *Aggregation*: BibSonomy aggregates tags in bags (as opposed to sets). Tags are shown in tag clouds where the size of a tag represents its frequency.
- *Object type*: The two resource types, web links and publication references, are both textual.
- *Source of Material*: Bookmarks and publication references are a mix of global resources and user-generated content. Web pages and publications are fix entities, more or less available to anyone. However, their representation in BibSonomy, particular for the case of publication references, is a product of user activity. Publications are identified through the metadata users enter. Users can choose what they input, they can create erroneous or incomplete entries or find different correct ways to fill a field (e.g., abbreviated journal names versus full names).

<sup>12</sup><http://www.bibsonomy.org/bibtex/13cd50bc064b9659829229f42eee284dd> (accessed June 2, 2015)

<sup>13</sup><http://www.kde.cs.uni-kassel.de/ws/dc09>

- *Resource connectivity*: Web pages can be connected by links. Publications can reference other publications. Furthermore, they can share the same authors, venues, publication year, and so on. In BibSonomy itself, publications are linked through their authors.
- *Social Connectivity*: In BibSonomy, it is possible to create (directed) links to other users by declaring friendships. The friendship can but need not be reciprocated. Users can restrict the visibility of their posts so that only their friends can see them. Furthermore, users can send posts to the inbox of users who have declared friendship. Moreover, users can create and join groups, to exchange documents, to make posts available to the group's members, or simply to be part of the group and thus to contribute to the group's representation as the union of its members personomies.

Next to the folksonomy-typical features, like tagging, navigation along the folksonomic entities (users, tags, and resources), or tag search, BibSonomy offers further features, like:

- Publications can be exported into various citation formats and styles.<sup>14</sup> The central format in the system is BIB<sub>T</sub>E<sub>X</sub>. Using Citation Style Language (CSL), lists of publications can be formatted using a huge variety of styles.
- To make posting easier, BibSonomy offers browser plugins. While visiting a website, by the click of a button, a user can post the website to BibSonomy, storing it as a link, or extracting publication metadata.
- BibSonomy offers an Application Programming Interface (API) allowing other software to automatically access the data for the purpose of integrating it into other services. BibSonomy can thus be used to compile publication lists for researchers' homepages or in content management systems, in e-learning software or L<sub>A</sub>T<sub>E</sub>X editors. At the time of writing, more than 20 tools are available.<sup>15</sup>
- Users may not only visit, copy, or export publications and bookmarks; they can also discuss these resources and even rate them.

BibSonomy has been the subject of various studies, of which we mention several in the next section. Datasets of BibSonomy are publicly available: The public part of the tagging data (all public posts with their user, tags, and resource) is made available in anonymized form to the community of researchers. The generation of these datasets is described in [Jäschke et al., 2012]. Similarly, BibSonomy's usage data (log files) is also made available upon request.<sup>16</sup>

---

<sup>14</sup><http://www.bibsonomy.org/export>

<sup>15</sup><https://bitbucket.org/bibsonomy/bibsonomy/wiki/Integration%20with%20other%20Websites%20and%20Services> (accessed June 2, 2015)

<sup>16</sup><http://www.kde.cs.uni-kassel.de/bibsonomy/dumps/>

### 2.3.3 Research Directions on Social Bookmarking

Lots of research has already been conducted in the field of social bookmarking. A comprehensive discussion of various aspects of social bookmarking has been presented by Peters [2009]. In the following, we review examples of previous studies in different directions, to demonstrate the high and diverse interest that these systems have gained from the research community. Due to the large amount of literature on tagging, we only mention examples from each topic. Since social bookmarking is one of the recurring topics in this thesis, more related work will be discussed in the respective chapters. Particularly in Chapter 5, we discuss four aspects of social bookmarking usage and we provide references to the literature for each aspect.

**Inception of Tagging.** Work on social tagging and emerging folksonomies began in late 2004, when the term *folksonomy* was coined by Vander Wal [2007], and it continued in 2005 in various blog posts and papers. One of the first reviews about social tagging systems was provided by Mathes [2004]. He noted that social tagging systems allow a much greater variability in organizing content than formal classification can provide. Mathes further identified some potentials and uses of tagging systems, such as serendipitous browsing. He was also among the first to hypothesize that tag distributions follow power-law distributions, which can characterize the semantic stabilization of such systems (see also Golder and Huberman [2006]). The nature of the distributions was confirmed by Cattuto et al. [2007], who also investigated a series of further properties of the folksonomy structure. Noy et al. [2008] used BibSonomy as one candidate in the Collaborative Knowledge Construction (CKC) Challenge with the goal to infer users' expectations towards different tools "for collaborative construction of structured knowledge". In that challenge, it was found that users share several expectations towards such systems, like a suitable web interface, private and public spaces, or export facilities. Several users expressed their wish that tagging should also be integrated into further services (other than specific tagging systems).

**Emergent Semantics.** Following the idea that the way in which users employ tags might tell us something about these tags' semantics, Cattuto et al. [2008] and Benz et al. [2010b] analyzed the potential of emergent semantics in folksonomies. They showed, that folksonomy data can be used to discover synonyms or hierarchical relations. Körner et al. [2010] used tagging behavior to differentiate between two types of users, "categorizers" and "describers", and found that the contribution of describers, who add more descriptive tags to their resources than categorizers, benefits the mining of semantics.

**Information Retrieval in Tagging Systems.** Golder and Huberman [2006] studied the tags themselves and identified seven different kinds of tags depending on the function they have within the collection of bookmarks. Peters and Weller [2008] summarize a variety of activities for manipulating the set of tags in a folksonomy or

their presentation in tag clouds to improve their usefulness for retrieving resources. Moreover, Peters and Stock [2010] and Peters et al. [2012] proposed computing “power tags” per document, that is, the most frequently used tags for a document, cut-off at some threshold. A special tag search that retrieves only those documents where the queried tags are power tags, is particularly successful for one-tag queries. The task of ranking resources was, for example, approached by [Hotho et al., 2006c], who introduced the *FolkRank* algorithm, which was later also used to produce recommendations and which will be a topic in Part III of this thesis.

**Folksonomic Recommendation.** A large body of literature exists on recommending any of the folksonomic entities in tagging systems. Most of it has focused on recommending tags, most noticeable the two tag recommender challenges at ECML/PKDD 2008 [Hotho et al., 2008] and 2009 [Eisterlehner et al., 2009]. Others have tackled the tasks of recommending resources [Bogers, 2009] or users [Manca et al., 2015]. We go into further detail in the next section. Apart from the development and evaluation of suitable recommendation algorithms also their real-world impact on the tagging system has been studied, for example in terms of vocabulary breadth or quality [Font et al., 2016]. More details on folksonomic recommender systems follow in Section 2.4.

**Scholarly Usage.** Borrego and Fry [2012] used the publicly visible publication posts of BibSonomy to analyze the use of publications by scholars. They found that in BibSonomy, articles that appeared in commercially published journals significantly outnumber the articles from other sources like science-minded repositories or open access journals. Haustein and Peters [2012] compared the use of tags for scholarly publications to other publication descriptions, like words from the title or from the abstract, or terms from automatic indexing. They found only small overlaps between tags and the other descriptions and concluded that the readers’ perspective (reflected in the tags) is indeed a valuable supplement to the perspective of authors and indexers. A lot of research has been spent on the system Mendeley and we discuss that as well as further literature on the connections between scholarly impact and usage in a social bookmarking in our study of that subject in Chapter 6.

**Further Topics.** Other studies have investigated multilingual tagging [Stiller et al., 2011], the consistency of community structures extracted from various interaction networks between users (evidence networks) [Mitzlaff et al., 2011], or spam detection [Krause et al., 2008].

## 2.4 Folksonomic Recommender Systems

Recommender systems have become a vital part of the web, where they assist users in their content selection by pointing to personalized sets of resources. In this thesis, recommender systems play a role in Chapters 7 and 8 and therefore, we briefly



present the field of recommender systems and particularly the typical settings for recommendations in social tagging systems. The classic scenario comprises a set of users, a set of items, and for some user-item pairs the knowledge whether or even how much the user liked the item. The goal of a recommender system is to suggest items that the user at hand (the *active* user) will find interesting but for which it is not explicitly known whether or not they actually like them. Typically, the available data is rather sparse, meaning that the set of items is large but only for very few items, a user's actual preference is known.

Recommender algorithms can roughly be classified into three classes: *Content-based* algorithms use the content of resources, for instance, to compute similarities between items and to present items that are similar to the ones the active user previously liked. *Collaborative* algorithms make use of the relations between the users and the items, for instance, by identifying similar users and suggesting items, similar users liked. The third class are algorithms that exploit both data sources, sometimes called *hybrid* recommenders.

Recommender systems exist for all kinds of resources, like movies, books, videos; and the study and development of recommender systems has become its own scientific discipline. A platform for many recent results on the topic is the RecSys conference series.<sup>17</sup> In this thesis, we focus on recommendations in tagging systems, thus, scenarios in which the folksonomy data is used to provide a user with recommendations for tags to use in a post or for resources they might find interesting.

### 2.4.1 Recommendations in Social Bookmarking Systems

Recommender systems in social bookmarking systems can provide recommendations for all three kinds of entities, users, tags, and resources. The employed folksonomic recommender algorithms can exploit the folksonomy structure, thus the relations between the entities, the folksonomy hypergraph, cooccurrences, popularity, and so on. In this thesis, we discuss recommendations for tags and resources as, other than recommending users, these two support the most common use case in a social bookmarking system: Users create a post, that is, they store a resource and annotate it with suitable tags.

*Tag recommenders* provide recommendations during the posting process. When a user  $u$  has decided to post a resource  $r$ , the recommender will provide a list of tags that  $u$  might want to assign to  $r$ . Note that we can neither assume  $u \in U$  nor  $r \in R$ , since both the user and the resource might be new to the system. Furthermore, the suggested tags may or may not be chosen from the set of all previously used tags  $T$ . New tags can, for example, be generated from the resource itself (e.g., using the words of a publication's title if the resources are publications). Recommending tags can serve various purposes, such as: increasing the chance of getting a resource annotated, reminding a user what a resource is about, and consolidating the vocabulary across

---

<sup>17</sup><http://recsys.acm.org/>

users. Furthermore, as Sood et al. [2007] pointed out, tag recommenders lower the effort of annotation by changing the process from a *generation* to a *recognition* task: rather than “inventing” tags, the user only needs to select some of the recommended tags.

*Resource recommenders* produce a personalized list of resources for an active user. They can be shown to the user upon request and should contain resources the user did not yet know but might find interesting.<sup>18</sup> Resource recommendations can help overcome the information overload problem users face in tagging systems. Although usually all content can be reached by navigating the folksonomy structure or through search, recommenders can be of help when there is simply too much content available. Recommenders can point the active users to resources they would otherwise not – or only with great effort – have found.

### 2.4.2 Evaluation of Folksonomic Recommendations

When new recommender algorithms are proposed, they must be compared to previous approaches to evaluate their performance. In Chapter 7, we scrutinize the typical evaluation setup, and in Chapter 8, we conduct benchmarking experiments comparing different recommendation strategies. Therefore, in the following, we discuss the evaluation of recommender algorithms in tagging systems.

Recommender systems can be evaluated in three ways: in an offline evaluation, in a user study, or online [Shani and Gunawardana, 2011]. All three methods have their benefits and drawbacks with regard to their cost, their biases, and their discrepancies to the actual productive application. Different methods may yield different results. For example, McNee et al. [2006] found differences in the performances of algorithms in offline evaluations and user studies indicating a gap between *accurately* predicting items a user will choose versus providing actually *useful* recommendations. However, due to the high costs of the alternatives, many recommender benchmarks are conducted as offline studies and we describe typical experimental setups in this section. In Chapter 7, we will investigate the setups in more detail and point to several pitfalls. Here, we describe the general procedure, in which a historic dataset is used to evaluate recommender algorithms in a prediction scenario.

#### Test and Training Data

In offline evaluations historic data serves as gold standard. A dataset is split into a *test* and a *training* set. The latter is used to train an algorithm. Instances from the test set are selected and those entities that are to be recommended (e.g., the tags or the resources) are removed. The remainder is used as input for the recommender

---

<sup>18</sup>There are other kinds of resource recommendations as well, for example, producing recommendations based on a user and a resource (similar items), or – similarly to search – based on a user and a query. Moreover, there are scenarios in which it is meaningful to recommend resources that the active user did use before, such as songs the user previously listened to for a play list. In this thesis, we consider, however, the task of recommending new resources, given only the active user.

algorithm, which produces a ranked list of recommendations for each entity. Each list is compared to the respective set of removed entities: Since they are the user’s actual choice, it is assumed that a good recommender would rank many of them high.

A commonly used method to produce training and test data is  $n$ -fold cross validation: The original dataset is partitioned into  $n$  folds (subsets). Then one set is selected as test set and the other  $n - 1$  sets for training. Thus, all experiments can be repeated  $n$  times. A variation of cross validation are hold-out procedures, which have become the dominant strategy in the offline evaluation of folksonomic recommendations. For folksonomy datasets, a procedure called *LeavePostOut* is often used to evaluate tag recommendations, while the variant *LeaveXPostsOut* is applied to test resource recommenders.

**LeavePostOut.** A variant of the leave-one-out hold-out estimation [Herlocker et al., 2004] is *LeavePostOut* [Jäschke et al., 2007]. Given a dataset, one experiment consists of the following steps: For each user  $u \in U$ :

1. One post  $p$  is selected at random.
2. The post  $p$  is eliminated from the dataset and the remaining data is used for training.
3. The task for the recommender algorithm then is to produce tag recommendations (i.e., to predict the tags of  $p$ ) given both the user and the resource of  $p$ , while the tags of  $p$  serve as gold-standard (and are therefore of course withheld from the algorithm).
4. A score is assigned that measures the prediction quality of the recommendation, comparing the list of recommended tags to the actual tags of post  $p$ .

This procedure is repeated for every user  $u \in U$  and the resulting scores are averaged. In a way, *LeavePostOut* can be considered a  $|U|$ -fold cross validation.

**LeaveXPostsOut.** A variation of *LeavePostOut* is *LeaveXPostsOut* where, in contrast to the former procedure, a set of several posts  $X_u$  of the same user  $u$  is left out at once (e.g., Bogers [2009]). The user  $u$  is then used as input for the recommendation algorithm. This is particularly useful for the item recommendation scenario, where for an active user  $u$ , a ranked list of recommended posts must be produced. In that case, leaving out only one post, would require the recommender to predict a single resource. Leaving out the set of posts  $X_u$  makes it possible that more than one post of the recommender’s result list will be considered relevant to user  $u$ . It is assumed that a good recommender would rank many of the withheld posts in  $X_u$  among the first positions in the list of recommendations. *LeaveXPostsOut* is repeated for every user and the resulting scores are averaged.

Compared to evaluation methods where the dataset is divided only once into a fixed training set and a fixed test set (used in traditional classifier evaluation), the

*LeavePostOut* and *LeaveXPostsOut* methods are unbiased by the selection of those users in the test set as each user is considered in the evaluation. This advantage is of importance especially on small datasets where one can not consider an arbitrarily chosen (small) sample of users to be representative for the whole dataset.

### Evaluation Measures

Both above evaluation scenarios produce for each user  $u \in U$  a list of recommended entities (tags or resources) and a list of actual entities (the actual tags of the left-out post, or the left-out resources) to compare to. In this thesis, we use the three evaluation measures recall, precision, and mean average precision, which can be found in [Manning et al., 2008], and which we describe here: For a user  $u$ , let  $E_u$  be the set of left-out entities and  $e_{u,i}$  be the entity that the recommender placed at position  $i$  in the list of recommendations for user  $u$ . Then the *recall at  $k$* ,  $\text{rec}@k$ , counts the share of entities in  $E_u$  that have been placed among the top  $k$  positions in the list of recommendations:

$$\text{rec}@k(E_u) := \frac{1}{|E_u|} |\{e_{u,1}, e_{u,2}, \dots, e_{u,k}\} \cap E_u| .$$

The *precision at  $k$* ,  $\text{pre}@k$ , counts the share of entities among the top  $k$  positions that belong to  $E_u$ :

$$\text{pre}@k(E_u) := \frac{1}{k} |\{e_{u,1}, e_{u,2}, \dots, e_{u,k}\} \cap E_u| .$$

Both measures produce one score for each user. To get an overall impression on an algorithm's performance, these values are averaged over the set of all users to yield the final.

The parameter  $k$  is called the *cut-off* level, since only the top  $k$  recommendations are evaluated. It must be chosen by the experimenter. For tag recommendations, a common choice is  $k = 5$  since proposing five tags per recommendation seems a suitable choice in a tagging system, where the recommendations are shown only during the process of posting, which usually takes only a short amount of time. However, it might well be that different users would prefer different numbers of recommended tags. For resource recommendation, there is no such typical choice. To avoid having to set  $k$ , one can use the *mean average precision* measure MAP, which is unparametrized: For each user  $u$ , the *average precision* AP is computed as

$$\text{AP} := \frac{1}{|E_u|} \sum_{k=1}^n \text{pre}@k(E_u) \cdot \delta(E_u, e_{u,k}) ,$$

where  $\delta(E_u, e_{u,k})$  indicates whether the resource ranked at position  $i$  is one of the withheld resources of the user  $u$ , and  $n$  is the length of the list of recommended items. The resulting MAP score for an algorithm is calculated as the mean of the average precision for each user.

### 2.4.3 Folksonomic Recommender Algorithms

In this section, we introduce the algorithms we apply in Part III of this thesis. Over time, many approaches have been proposed and we enumerate several in the respective sections on related work in Chapter 7 on tag recommender systems and in Chapter 8 on item recommendation. A recent survey by Godoy and Corbellini [2015] selected more than 130 papers on recommending tags, resources, or users in social tagging systems, grouping them by their methodology. In this thesis, we analyze the *FolkRank* algorithm in-depth in Chapter 8. All other algorithms are only used as baselines or are part of the evaluation of recommendation setups in Chapter 7, where we do not explore particular properties of the algorithms. Therefore, we content ourselves with short descriptions and refer for details to [Jäschke et al., 2008] and to Chapter 8 for *FolkRank* and *adapted PageRank*.

The most simple recommender is *most popular tags*, which is often used as a baseline in recommender benchmarks. It always produces the same recommendations: the most frequently used entities (resources or tags) in the system. It is thus an unpersonalized recommender that can be used for resource as well as for tag recommendations.

Almost as simple, but personalized and much more effective are the two tag recommenders *most popular tags by resource*, and *most popular tags by user*. Given the active user  $u$  and the resource  $r$  to be posted, the former recommends the tags that have previously most often been used for  $r$  (by other users), and the latter recommends the previously most often used tags of  $u$  (for other resources). Since in the resource recommendation scenario only the active user  $u$  is given as input, and since in contrast to tag recommendations, it would not be reasonable to suggest resources a user had previously already posted, there are no respective counter parts of these two algorithms for resource recommendation.

For resource recommendation, we will use *collaborative filtering* [Sarwar et al., 2001], which recommends new resources to an active user based on the preference of like-minded users. The algorithm has been invented for a scenario in which users explicitly rate items. Each user  $u$  is represented as a vector  $\vec{x}_u$ , where an entry  $\vec{x}_{u,i}$  for some item  $i$  describes that user's rating, if it is known. Based on these vectors, a set of the  $k$  most similar users is computed, and recommended are items from the collections of these users. They are ranked according to the collective ratings aggregated over these  $k$  users. Since the folksonomy data does not contain explicit user ratings for resources, we interpret the fact that a user bookmarked a resource as (Boolean) expression of the user's interest in that resource. To this end, we reduce the ternary relation  $Y$  of the folksonomy  $\mathbb{F}$  to a lower dimensional space as described by Jäschke et al. [2008]. The vector  $\vec{x}_u^R \in \mathbb{N}^{|R|}$  represents the number of tags that user  $u$  has assigned to resources  $r \in R$ : For each  $r \in R$ , we set  $\vec{x}_{u,r}^R := |\{t \in T \mid (u, t, r) \in Y\}|$ . Alternatively, we represent users by the tags they have used with a vector  $\vec{x}_u^T \in \mathbb{N}^{|T|}$ : For each tag  $t \in T$ , we set  $\vec{x}_{u,t}^T := |\{r \in R \mid (u, t, r) \in Y\}|$ , the number of resources, this tag has been applied to by user  $u$ . This variant is called  $CF_T$  in the sequel, the resource-minded one is called  $CF_R$ . The parameters of *collaborative filtering* are the number of similar

users  $k$  and the similarity function. Furthermore, one can reduce the vectors  $\vec{x}_u^R$  or  $\vec{x}_u^T$  to Boolean versions: For example, in the Boolean version of  $\vec{x}_u^R$ , each entry  $\vec{x}_{u,r}^R$  equals  $\vec{x}_{u,r}^R = 1$  if user  $u$  has bookmarked  $r$ , and  $\vec{x}_{u,r}^R = 0$  otherwise.

The *adapted PageRank* and *FolkRank* were first presented in [Hotho et al., 2006c]. Both are adaptations of the original *PageRank* algorithm [Brin and Page, 1998] to the ternary hypergraph of folksonomies. Both are described in Section 8.2.

## 2.5 German Laws with Relevance for Rating and Reviewing

With the advent of the Web 2.0, a variety of forums for the (public) utterance of own opinions has been established. Blogs, wikis, reviewing systems, and so on, present the opportunity to voice one's own opinion to a large public, basically to anyone who has access to the web, often anonymously. For scholarly literature such discussion has become known as social peer review or post-publication peer review. Among others, social bookmarking systems (e.g., BibSonomy or CiteULike) have been extended with respective features, allowing their users to rate and comment on scholarly publications.

As a consequence, the question of what may or may not be published in such forums has arisen and with it the question of who is responsible for the content. The answers to these questions are non-trivial and they have been discussed controversially in the past. As these are questions of law, they are beyond the scope of this thesis. The contribution in this work (Chapter 9) is rather a discussion of technical design options in Web 2.0 systems, aiming to minimize risks for those who operate or use such systems as well as for those who become subject of opinions that are published in these systems. For more details on the judicial foundations, we refer to [Doerfel et al., 2013b].

In this section, we briefly mention fundamental rights (from the German legal system<sup>19</sup>) that need to be considered when dealing with web systems that are likely to contain opinions about persons, like reviewing systems (in this thesis particularly relevant for scholarly literature). The English translation of several articles of German Law are taken from the English language version provided by the German Federal Ministry of Justice and Consumer Protection.<sup>20</sup>

The first sentence of Article 5 of the Basic Law for the Federal Republic of Germany<sup>21</sup> states: “*Every person shall have the right freely to express and disseminate his opinions in speech, writing and pictures, and to inform himself without hindrance from generally accessible sources.*” It is the basis for two important rights: the freedom of opinion and the freedom of information. These two rights protect the one who utters an opinion, and its recipient, the one who reads or hears it. The one who the opinion is about is

---

<sup>19</sup>Of course, legal requirements are different in other countries. The particular laws are however not the central topic in this work. German law was chosen since the author is German and since the web system BibSonomy, that provides the use case in several parts of this thesis, is hosted and operated in Germany.

<sup>20</sup><http://www.gesetze-im-internet.de/>

<sup>21</sup>[http://www.gesetze-im-internet.de/englisch\\_gg/englisch\\_gg.html](http://www.gesetze-im-internet.de/englisch_gg/englisch_gg.html) (accessed October 10, 2015)

protected by the personality rights. They are anchored in German Basic Law with the first paragraph of Article 1: *“Human dignity shall be inviolable. To respect and protect it shall be the duty of all state authority.”* and the first paragraph of Article 2: *“Every person shall have the right to free development of his personality insofar as he does not violate the rights of others or offend against the constitutional order or the moral law.”* In German law, personality rights are structured along different levels of protection (called spheres). An opinion about a person can touch this person’s private sphere<sup>22</sup> or the less protected social sphere.<sup>23</sup>

Between these fundamental protective laws (freedom of opinion, freedom of information, and personality rights), an area of conflict arises, as it has to be determined which law weighs stronger. For instance, when an opinion about a person is expressed publicly, it might be an intrusion into that person’s personality rights. However, depending on the content of the opinion, the protection of the opinion or the valid public interest might weigh stronger than the personality right, in which case the person must bear the public statement.

When information or opinions about scientific work are concerned – like in social peer review, which we discuss in Chapter 9 – the freedom of science is an important basic right. It has its legal basis in Article 5 of the Basic Law for the Federal Republic of Germany where it says: *“Arts and sciences, research and teaching shall be free.”*

All online systems that collect data from or about users are subject to privacy law and must respect the individual’s right to informational self-determination, the right to decide which information about oneself are divulged and processed. The particularities of this form of protection are regulated in the German law system by the Telemediengesetz (TMG)<sup>24</sup> and the Federal Data Protection Act.<sup>25</sup> An aspect that is of particular importance for the collection of user data, is the principle of data reduction and data economy (Section 3a, BDSG):<sup>26</sup> *“Personal data are to be collected, processed and used, and processing systems are to be designed in accordance with the aim of collecting, processing and using as little personal data as possible. In particular, personal data are to be aliased or rendered anonymous as far as possible and the effort involved is reasonable in relation to the desired level of protection.”*

---

<sup>22</sup>German: Privatsphäre

<sup>23</sup>German: Sozialsphäre

<sup>24</sup>The German term “Telemedien” covers electronic information and communication services (Article 1, TMG).

<sup>25</sup>Bundesdatenschutzgesetz (BDSG).

<sup>26</sup>[http://www.gesetze-im-internet.de/englisch\\_bdsch/englisch\\_bdsch.html](http://www.gesetze-im-internet.de/englisch_bdsch/englisch_bdsch.html) (accessed October 10, 2015)





## **Part I**

# **Analyzing Research Communities in Conferences**



## Chapter 3

### Analyzing a Community through its Conferences



In this chapter, we investigate a research community through various analyses of metadata, gathered from that community’s conference publications – thus, data from the creation phase of the publication life cycle. For that purpose, we choose the community of formal concept analysis (FCA). FCA was introduced by Rudolf Wille as a means of qualitative data analysis and visualization. The methodology has since been extended and now also provides means for quantitative analysis. We propose and demonstrate the use of FCA methodology and the typical FCA visualizations for analyzing corpora of scholarly literature. Such analyses benefit researchers in the respective community or those who plan to join it. For demonstration purposes we apply the FCA techniques to a dataset comprising all publications in the history (up to 2011) of the three conferences most relevant to the FCA community. Using FCA, we investigate patterns and communities among authors, we identify and visualize influential publications and authors, and we give an overview on the distributions of community members over the conferences. We complement the analysis with typical statistics from bibliometrics. This chapter is an extension of our work in [Doerfel et al., 2012b].

### 3.1 Introduction

This chapter commences the publication life cycle theme of the thesis. We start with the analysis of a research community through its conferences – particularly, through metadata of the conference proceedings, thus data from the creation phase of the scholarly publication life cycle. We aim at answering typical questions that every researcher faces from time to time, such as: (i) Which are the most influential authors, papers, and conferences in that area? (ii) Who is cooperating with whom on which topics? (iii) Who is citing whom? It can be tedious to keep an overall overview about one’s general area of research, even for long-term researchers. Analyses of a research community allow its members to gauge and to extend their perception of that community. They also can be of help for newcomers and allow faster access to the

community by pointing to the most influential authors or to the must-read literature of a particular subtopic within a larger area of research.

In this chapter, we propose the use of formal concept analysis (FCA) as a tool for the investigation and visualization of different aspects of a scientific community. FCA provides a set of methods for qualitative data analysis and visualization. Such qualitative analyses are suitable to accompany the traditional measures from bibliometrics (like publication or citations counts or impact factors), thus creating a more whole picture of the community: While they do not yield a measure for individuals (papers, authors, or venues), that would allow comparing these entities among each other, they rather identify relations between them and even between different kinds of entities, like between authors and conferences or between authors and publications.

For this study, we select the FCA community itself as our subject of investigation. Our intention for this analysis is to gain more insights into the structure of the FCA community and its relationship to closely related disciplines. For that purpose, we create a dataset containing the metadata of all publications of the three conferences series, most relevant to the FCA community: ICCS, ICFCA, and CLA. To the resulting corpus of metadata, we apply FCA methodology next to other, more traditional means of bibliometrics. In our analysis, we target two different levels: the conference level and the author level.

**Research Questions.** In this chapter, we introduce the methodology of FCA, accompanying traditional quantitative bibliometric analyses, to provide answers to the following questions:

- (RQ1) How do the different conference series or individual editions compare in terms of impact and participation?
- (RQ2) Who collaborates with whom and who has been influential to the individual groups?
- (RQ3) Who are the key players within the community and which roles do they assume?

Providing answers to these questions enables researchers to get an overview about their community and to keep track of it. While in this chapter, we discuss the above questions for the FCA community, the methodology can easily be transferred to other communities and their conferences. Apart from the interpretation of the results, all here presented approaches can be automated and applied to produce and update analyses of other communities as well. The application to the FCA community is an example that demonstrates the kinds of analyses by which researchers of a community can be supported.

**Contributions.** In the following, we exploit data from the creation phase of the publication life cycle to present a comprehensive study of the FCA community:

1. We introduce FCA methodology to the analysis of publication metadata, to discover influential authors and publications, and to investigate conference participation. We make our findings explicit using the typical FCA visualization as line diagrams.
2. We demonstrate this methodology on the example of the FCA community.
3. We accompany these rather qualitative results with typical bibliographical statistics and thus present a comprehensive study of the FCA community – observed through its three most relevant conference series.
4. The dataset of this study – an intensely cleaned corpus containing metadata of all publications in the three conferences series (until 2011) that are the most relevant to FCA – is made publicly available.<sup>1</sup>

**Limitations.** Since the focus of our analysis is on the three conference series, many publications related to FCA (in particular journal articles, textbooks, and publications before 1993, and thus the fundamental articles introducing FCA) have not been included in the dataset.<sup>2</sup> Our analysis therefore is never an evaluation – measures from bibliometrics sometimes are – of authors’ contributions in general or to FCA specifically. Instead, we present an overview and interesting connections between authors and publications of the three conference series that are the main forum for the current advances in the field.

The means of analyzing a scientific community are demonstrated using the example of the FCA conferences. However, the applied methods, particularly the application of FCA to publication metadata, can similarly be applied to any other scientific community. The only requirement is the availability of a similar dataset.

**Structure.** The structure of this chapter is as follows: In the next section, we briefly introduce the basics of FCA. Afterwards, in Section 3.3, we discuss related work, and in Section 3.4, we describe the dataset of publications in detail. Section 3.5 provides the results and demonstrates the use of FCA in the analysis of publication metadata. Finally, in Section 3.6, we conclude the chapter and address future work.

In various analyses in the following, we mention individual publications from our corpus or their references. These references are not included in the reference section at the end of the thesis but can be found in an extra reference list in Appendix A.

Several parts of the study in this chapter have previously been published in [Doerfel et al., 2012b]. These parts have been extended (the citation and impact analysis in Section 3.5.1 and the discovery of community roles in Section 3.5.2) and updated (citation counts in Table 3.2). All parts have been rearranged for this thesis. [Doerfel

---

<sup>1</sup><http://www.kde.cs.uni-kassel.de/datasets/>

<sup>2</sup>These publications can of course be referenced by publications in the corpus and thus occur in those analyses that include references, for example, in Section 3.5.2.

et al., 2012b] contains additional analyses (on author fluctuation and on the most often cited publications). These have been excluded from this thesis as they had been contributed by co-author Robert Jäschke.

## 3.2 Formal Concept Analysis

The definition of *formal concepts* (within a *formal context*) is a mathematical approach to define the notion of concepts and of hierarchical relations among them. Wille [1982] based his definition of a formal concept on foundations from philosophy, making a concept essentially a pair of two things: (i) *the extent* (i.e., the objects that belong to the concept) and (ii) *the intent* (i.e., the properties that these objects have in common). Wille laid out the basic foundations for what became the field of *formal concept analysis* (FCA) as “an attempt to unfold lattice-theoretical concepts, results, and methods in a continuous relationship with their surroundings” [Wille, 1982]. FCA has thus become a field of applied mathematics with strong roots in mathematical algebra – specifically (complete) lattice theory. A mathematical abstraction of conceptual thinking is the foundation of the theory, and it has been the basis for many practical applications – most prominently for visualizing conceptual hierarchies as easily interpretable line diagrams.

In the following, we recall some of the basic definitions of FCA. They can be found similar in [Wille, 1982]. However, since the textbook by Ganter and Wille [1999] has become the definitive book on the mathematical foundations FCA, we use it for reference in the remainder of this section. A broader introduction to lattices and FCA can be found in Chapters 0 and 1 of that book. We begin with the definition of a *formal context* as an abstraction of objects and their properties.

**Definition 3.1** (Formal Context, cf. Definition 18 in [Ganter and Wille, 1999]). A formal context is a triple  $\mathbb{K} := (G, M, I)$ , where  $G$  is a set of objects,  $M$  is a set of attributes,<sup>3</sup>  $I \subseteq G \times M$  is a binary relation between  $G$  and  $M$ . With  $(g, m) \in I$ , we denote that “the object  $g$  has the attribute  $m$ ”.

Within such a (formal) context, we can now identify concepts:

**Definition 3.2** (Formal Concept, Extent, Intent, cf. Definitions 19 and 20 in [Ganter and Wille, 1999]). A formal concept of a context  $(G, M, I)$  is a pair  $(A, B)$  with  $A \subseteq G$ ,  $B \subseteq M$  and

$$B = A' := \{m \in M \mid \forall g \in A: (g, m) \in I\}$$

and

$$A = B' := \{g \in G \mid \forall m \in B: (g, m) \in I\}.$$

The set  $A$  is called the *extent* and the set  $B$  is called the *intent* of the formal concept  $(A, B)$ .

---

<sup>3</sup>The letters  $G$  and  $M$  stand for the respective German words for objects and attributes: “Gegenstände” and “Merkmale”.

Simply put, a pair  $(A, B)$ , with  $A$  being a set of objects and  $B$  being a set of attributes, is a formal concept if  $A$  contains exactly all those objects that share all the attributes in  $B$ , and  $B$  contains exactly all the common properties of the objects in  $A$ . The set of concepts can be ordered by the sub-set relation  $\subseteq$  on their extents (or equivalently by the super-set relation  $\supseteq$  on the intents). Furthermore, it is easy to see (compare [Ganter and Wille, 1999], Proposition 11) that for any set  $X$  of concepts of a context  $\mathbb{K}$ , the intersection of their extents always yields an extent of  $\mathbb{K}$ , and the intersection of their intents always yields an intent of  $\mathbb{K}$ . It follows for any such set  $X$  that there always is a largest concept that is smaller than every concept in  $X$ , called the *infimum* of  $X$ . Similarly, there always is a smallest concept that is larger than every concept in  $X$ , called the *supremum* of  $X$ . Thus, the set of all concepts of a context always forms a complete lattice:

**Definition 3.3** (Complete Lattice, cf. Definitions 9 and 10 in [Ganter and Wille, 1999]). *An ordered set  $V := (V, \leq)$  is a complete lattice if for any subset  $X \subseteq V$*

- *its supremum  $\bigvee X$  – an element of  $V$  such that  $\forall x \in X : x \leq \bigvee X$  and for all elements  $v \in V$  with the same property holds  $\bigvee X \leq v$  – and*
- *its infimum  $\bigwedge X$  – an element of  $V$  such that  $\forall x \in X : \bigwedge X \leq x$  and for all elements  $v \in V$  with the same property holds  $v \leq \bigwedge X$  –*

*exist.*

**Definition 3.4** (Concept Lattice, cf. Definition 21 in [Ganter and Wille, 1999]). *The set  $\mathfrak{B}(\mathbb{K})$  of all formal concepts of a formal context  $\mathbb{K}$  together with the partial order  $(A_1, B_1) \leq (A_2, B_2) :\Leftrightarrow A_1 \subseteq A_2$  (which is equivalent to  $B_1 \supseteq B_2$ ) for concepts  $(A_1, B_1)$  and  $(A_2, B_2)$  is a complete lattice, called the concept lattice of  $\mathbb{K}$ .*

The concept lattice is a hierarchical structure in which concepts are ordered. A concept lattice can be visualized as a line diagram, where each concept is represented by a node. Two concepts  $(A_1, B_1) \neq (A_2, B_2)$  are connected by a line if  $(A_1, B_1) \leq (A_2, B_2)$ , and there is no concept 'between them':  $(A_1, B_1) \leq (A_3, B_3) \leq (A_2, B_2) \implies (A_1, B_1) = (A_3, B_3)$  or  $(A_3, B_3) = (A_2, B_2)$ . The concepts' nodes are laid out such that  $(A_1, B_1)$  is below  $(A_2, B_2)$ , and thus the connecting line ascends from  $(A_1, B_1)$ . In the diagram, each object  $g$  is annotated below the smallest concept that has  $g$  in its extent, and each attribute  $m$  is annotated above the largest concept that has  $m$  in its intent. Examples of concept lattice diagrams can be found here in Section 3.5.

Formal Contexts represent binary relations between objects and attributes, and thus, for each pair  $g \in G$  and  $m \in M$  of a context, we either have  $(g, m) \in I$  or  $(g, m) \notin I$ . A generalization of this model are *many-valued contexts*:

**Definition 3.5** (Many-Valued Formal Context, cf. Definition 27 in [Ganter and Wille, 1999]). *A many-valued context is a quadruple  $(G, M, W, I)$ , where  $G$  is a set of objects,  $M$  is a set of attributes,  $W$  is a set of values,<sup>4</sup> and  $I \subseteq G \times M \times W$  is ternary relation with the additional requirement:  $(g, m, w), (g, m, v) \in I \implies w = v$ .*

<sup>4</sup>The letter  $W$  stands for the German word for values: "Werte".

We say that an attribute  $m$  has the value  $w$  for the object  $g$  iff  $(g, m, w) \in I$ , and we denote that with  $m(g) = w$ . Using scales, one can derive a single-valued formal context from a many-valued formal context. Scales are themselves formal contexts that translate the values of an attribute into one or more (new) binary attributes. Section 1.3 in [Ganter and Wille, 1999] gives various examples for typical scales. In this chapter, we will use scaling to create single-valued contexts, relating authors and their visits to conferences in Section 3.5.1, and relating clusters of co-authors to most influential publications or authors in Section 3.5.2.

For any context  $\mathbb{K} = (G, M, I)$ , one can deduce implications between attributes. These are pairs of subsets  $A, B \subseteq M$  such that for all objects  $g \in G$  holds: If  $g$  has all the attributes in  $A$ , then it has also all the attributes in  $B$ , or simply speaking: from  $A$  follows  $B$ . Equivalent is the condition  $B \subseteq A''$ . Guigues and Duquenne [1986] constructed for each finite context a *basis* of attribute implications, which is a set  $\mathcal{L}$  of implications such that (i) every implication holding in the context can be derived from  $\mathcal{L}$ , and (ii) no implication in  $\mathcal{L}$  can be derived from the other implications in  $\mathcal{L}$ . This base is often referred to as *stem-base* or *Duquenne-Guigues basis*. For details see Section 2.3 of [Ganter and Wille, 1999]. We will use this stem-base in Section 3.5.1 to explore publication habits of authors.

In Section 3.5.2, we discuss the extents of an *iceberg lattice* of a context. An iceberg lattice is an ordered subset of the concept lattice, containing only concepts with extents larger (with respect to cardinality) than a given threshold (minimum support). Iceberg lattices and an efficient computation algorithm, called Titanic, have been introduced by Stumme et al. [2002].

## 3.3 Related Work

In this chapter, we review literature on FCA and on its previous application to the analysis of publication metadata. For related work on bibliometrics in general, we refer to Section 2.2.

### 3.3.1 Formal Concept Analysis

FCA has been established around 1980 by Rudolf Wille in Darmstadt, Germany. The basic definitions, as well as fundamental connections to lattice theory have been laid out and proven by Wille [1982]. Since then the mathematical framework has been further developed, and thus today, there is now a rich mathematical theory exploring various properties of contexts and concept lattices, their substructures, decompositions, factorizations, and more, as well as construction algorithms, for example to compute the set of all concepts of a context, like the *Next-Closure* algorithm from Ganter [1984]. The textbook by Ganter and Wille [1999] systematically presents the mathematical foundations of FCA, ranging from the basics to advanced algebraic theory. The visualization technique of conceptual structures as lattice diagrams, as well as various extensions of FCA have been applied in a diverse plethora of areas,



among others e-learning [Agrawal et al., 2014], neuroscience [Endres et al., 2012], gene analysis [González Calabozo et al., 2011], or a virtual museum [Wray and Eklund, 2011]. Many more examples for applied FCA can be found in the conference proceedings of the three conference series that are also the subject of this chapter: ICCS, ICFCA, and CLA. A comprehensive overview of more than one thousand papers from FCA and related disciplines is presented in the two part survey by Poelmans et al. [2013a,b].

### 3.3.2 Formal Concept Analysis of Publications

FCA has previously been applied to data on scientific publications. In the following, we review such papers, and we discuss similarities and distinctions to the work in this chapter. Tilley et al. [2005] presented a literature survey of 47 scientific publications in which FCA had been applied for software engineering. Next to summarizing and surveying the main contributions of the papers, the authors conducted a qualitative analysis of the paper corpus by visualizing concept lattices of different contexts that have the 47 papers as their object set. In four contexts, they related the papers to their topics – represented as software engineering activities and as the (number of) lines of code of applications described in the paper –, as well as to their authors and cited papers (from within the corpus). The latter two contexts were created solely from publication metadata, and thus that part of [Tilley et al., 2005] has the greatest similarity to our work here. While their visualization is suitable for the relatively small corpus of publications, it would become tedious to use similar visualizations for larger corpora, like the one we handle here (954 publications, for more details see Section 3.4). Therefore in this work, we use clustering techniques to detect communities of co-authorship, and we analyze them and their relations to the influencing papers and authors. Furthermore, we additionally consider venues, particularly three relevant conference series, and focus more on a description of the full FCA community rather than on a specific subtopic like software engineering in [Tilley et al., 2005].

Petersen and Heinrich [2008] presented several concept lattices relating publications to their authors and to citing publications. In a large lattice comprising publications of one particularly prolific author, they use ordinal scaling (on the number of citations) to zoom in on highly cited publications and their co-authors.

Poelmans et al. combined text mining and FCA to provide surveys on the FCA literature related to knowledge discovery [Poelmans et al., 2010], including 140 publications, and to information retrieval [Poelmans et al., 2012], including 103 publications. Recently, a two-part survey on FCA literature in general [Poelmans et al., 2013a,b] covered 1,072 publications. Next to a traditional survey of papers, they presented the relations of papers to their topics (determined using a thesaurus of relevant terms and text classification) in the line diagram of a respective formal context.

An early practical application of FCA to the management of literature has been presented by Rock and Wille [1997], who used metadata of publications to search and visualize a given publication corpus. Ferré [2007] created an algorithm to automatically

extract relevant sub-strings from given sets of strings. As a practical example, he presented a navigation tool for a corpus of all ICCS contributions published until 2005.

In contrast to all previous papers, we neither focus on a detailed analysis of a small publication corpus, nor on a rough statistical analysis of a large scale corpus. The medium size of our dataset (954 publications with 17,121 references) still allows us to look at specific authors or publications. We provide the first analysis of the three conference series, in particular the first analysis with a focus on FCA, that is applied next to such diverse methods as graph partitioning and ranking.

## 3.4 Dataset

For our analysis, we had to select a fixed corpus of publications. Since there is no comprehensive catalog of all FCA-relevant publications, we had to build the corpus ourselves. Research on FCA has been published in diverse venues (journals, conferences, workshops, and technical reports or even preprints), and thus it is virtually impossible to gather all FCA-relevant publications. Furthermore, for each paper that seems related to FCA, one would have to decide whether it truly is FCA-relevant or not (e.g., papers just mentioning FCA in their related work). However, that would require accessing these papers' content, which often is unavailable, and which constitutes an enormous effort considering the number of possible candidates.

We decided to focus on the three international conference series that constitute the main forum for FCA-related research: the *International Conference on Formal Concept Analysis (ICFCA)*, the *International Conference on Conceptual Structures (ICCS)*, and the conference *Concept Lattices and their Applications (CLA)*. ICCS began as a conference on Conceptual Graphs (CG), with first FCA papers in 1995, and a balanced contribution of CG and FCA papers a few years later; while both ICFCA and CLA focus on FCA topics.

Our corpus comprises the contributions to these conference series until 2011. Thus, our analyses target the active community that took part in these conferences and contributed to them. The focus on these three venues has also practical advantages: most of the publication metadata is available from the publishers, and accessing and processing the paper content is not required.

In the remainder of this section, we first describe how we collected the publication corpus, and then we define the data structures upon which our analyses are based.

### 3.4.1 Gathering and Preprocessing

We have gathered metadata for all papers published at any of the past editions (up to 2011) of the three conference series ICCS, ICFCA, and CLA: 19 editions of ICCS, 9 editions of ICFCA, and 7 editions of CLA,<sup>5</sup> see Table 3.1. Of each paper, we collected

---

<sup>5</sup>The first edition of the CLA 2002 in Horní Bečva was a small seminar with four talks, and hence, no published proceedings exist.

Table 3.1: Venues of the three conference series ICCS, ICFCA, and CLA.

conference	years and venues
ICCS	1993: Quebec City (CA), 1994: College Park (US), 1995: Santa Cruz (US), 1996: Sydney (AU), 1997: Seattle (US), 1998: Montpellier (FR), 1999: Blacksburg (US), 2000: Darmstadt (DE), 2001: Stanford (US), 2002: Borovets (BG), 2003: Dresden (DE), 2004: Huntsville (US), 2005: Kassel (DE), 2006: Aalborg (DK), 2007: Sheffield (UK), 2008: Toulouse (FR), 2009: Moscow (RU), 2010: Kuching (MY), 2011: Derby (UK)
ICFCA	2003: Darmstadt (DE), 2004: Sydney (AU), 2005: Lens (FR), 2006: Dresden (DE), 2007: Clermont-Ferrand (FR), 2008: Montreal (CA), 2009: Darmstadt (DE), 2010: Agadir (MA), 2011: Nicosia (CY)
CLA	2004: Ostrava (CZ), 2005: Olomouc (CZ), 2006: Hammamet (TN), 2007: Montpellier (FR), 2008: Olomouc (CZ), 2010: Sevilla (ES), 2011: Nancy (FR)

its title, its authors, and the cited references from the publisher website SpringerLink<sup>6</sup> (ICCS and ICFCA) or extracted these data from the paper’s PDFs found on CLA’s website.<sup>7</sup> In our dataset, invited talks, regular and short papers are treated the same; poster sessions, satellite workshops, as well as separate contributions proceedings were not considered.

Our preprocessing included the extraction of authors, titles, years, and references from HTML and PDF files using regular expressions and manual work. Further, we implemented several normalization and completion steps for the titles and author names, allowing matching and duplicate detection, and an extensive manual error correction. Therefore, we employed the normalization steps described in [Voss et al., 2009] with an additional removal of diacritics (e.g., ‘ä’ and ‘á’ were replaced by ‘a’). We used different heuristics – for example the Levenshtein distance [Levenshtein, 1966] – to find errors in author names and titles. All references without authors (often encountered for cited web pages) were removed from the dataset.

Since many publications were cited as different editions or prior to their publication (‘to appear’), we normalized the publication year by dating back different editions to the earliest mentioned date of publication. For example, the collected papers of Charles S. Peirce were cited with publication years 1931, 1935, 1953, 1958, and 1966, which we normalized to 1931: [Peirce, 1931].

<sup>6</sup><http://www.springerlink.com/>

<sup>7</sup><http://cla.inf.upol.cz/papers.html>

For the first ICFCA, 2003 in Darmstadt, no proceedings were published. Thus, we used the book by Ganter et al. [2005], which contains contributions from the participants of the first ICFCA on the state of the art of FCA and its applications.

To gain knowledge about publications outside of the corpus that cite any of the conference papers in our corpus, we retrieved citations from Microsoft Academic Search<sup>8</sup>. Note that these citations only roughly reflect the real number of citations a publication received since this search engine relies on citation data that is available on the web, and can remove errors and correctly match different citation variants only to a certain extent.

### 3.4.2 Notations and Derived Data Structures

We derived several structures (graphs and formal contexts) from the collected data that are described in detail in the following. All structures that use the references were created after removing self-citations (cited publications where one of the authors is also an author of the citing paper).

We denote the set of all authors that published at any of the three conferences by  $A$  and the set of all papers published at any of the conferences by  $P$ .

**Conferences.** To analyze the distribution of all authors over the three conference series, we use  $\mathbb{K}_{\text{conf}}^f = (A, \{ICCS, ICFCA, CLA\}, \mathbb{N}, I_{\text{conf}}^f)$ , a many-valued context where  $(a, c, n) \in I_{\text{conf}}^f$  iff  $a$  published exactly  $n$  papers at conference  $c$ . In other words,  $c(a)$  is the publication frequency of author  $a$  at conference  $c$ .

**Authorship.** The formal context  $\mathbb{K}_{\text{pa}} = (P, A, I_{\text{pa}})$ , with  $(p, a) \in I_{\text{pa}}$  iff  $a$  is an author of paper  $p$ , describes who authored which publication. The graph of co-authorship  $\mathfrak{G}_{\text{coa}}$  is an undirected, weighted graph with  $A$  as node set. Two authors are connected iff they published together, and their edge's weight is the number of co-authored publications at the conferences.

In Section 3.5.2, we cluster (partition)  $\mathfrak{G}_{\text{coa}}$ , or more specifically, its node set, to yield a community allocation for the authors (cf. Section 2.1.2). To create such a partitioning and its visualization in Figure 3.5, we laid out the graph using the force directed graph visualization provided by Graphviz [Gansner and North, 2000]. Then the *GMap* algorithm (again Graphviz), based on [Newman, 2006], was applied to discover communities of collaborators. GMap optimizes its output clustering with respect to *modularity* (cf. Definition 2.6), which is a community quality measure that compares the number of co-author edges within each community to the expected value for this number in an equivalent random graph. Finally, *Voronoi diagrams* are used to draw the ‘borders’ between the different communities, creating the look of ‘countries on a map’. We use these communities (clusters) as attributes of formal contexts. We denote by  $C_k(\mathfrak{G}_{\text{coa}})$  the set containing the  $k$  clusters with the highest cardinality.

---

<sup>8</sup><http://academic.research.microsoft.com/> (accessed December 2, 2014)

**Citations.** The directed, weighted graph  $\mathfrak{G}_{\text{cit}}$  has the authors in  $A$  as nodes. An edge  $(a, b)$  with weight  $w$  indicates that in all considered publications,  $w$  times, some publication of author  $b$  was referenced by author  $a$ .

To analyze influences, in Section 3.5.2, we relate co-authorship communities (see above) to publications and to authors that were cited by a community's members. This is modeled in the two many-valued contexts  $\mathbb{K}_{\text{cp}}^f = (C_k, P, \mathbb{N}, I_{\text{cp}}^f)$ , relating communities and cited publications, and  $\mathbb{K}_{\text{ca}}^f = (C_k, A, \mathbb{N}, I_{\text{ca}}^f)$ , relating communities and cited authors. For a cluster  $c \in C_k$  and a publication  $p \in P$ , we have  $(c, p, n) \in I_{\text{cp}}^f$ , or simply  $p(c) = n$ , iff publication  $p$  was cited exactly  $n$  times in publications by authors in  $c$ . Similarly, for an author  $a \in A$ , we have  $(c, a, n) \in I_{\text{ca}}^f$ , or simply  $a(c) = n$ , iff  $a$  was cited exactly  $n$  times in publications by authors in  $c$ . The parameter  $k$  controls the number of objects (i.e., co-authorship communities).

## 3.5 Analysis

In this section, we present the results of our analysis along two levels of entities: first on the level of conferences in Section 3.5.1 and then on the level of authors in Section 3.5.2.

### 3.5.1 Conferences

With the analysis of the three conference series, ICCS, ICFCA, and CLA, we approach our first research question (RQ1 from Section 3.1). We focus on two aspects: on the impact of the conference series and of individual editions – in terms of citations to and by the conference contributions – and on publication habits of the authors. For the first aspect, we use classic measures from bibliometrics. The second aspect can be tackled using FCA, which we use to identify behavior patterns in the visits of conferences. We start the section on conferences with some basic statistics in Table 3.2, that give an overview of the conference history and thus of the data in the corpus. We can see that, with its longer run-time and slightly more publications per conference, the ICCS series outnumbers the other two conference series in both authors and publications.

#### Citations from and to the Proceedings

The second block of Table 3.2 shows the statistics of citations by the publications in our corpus – found in the reference sections of the papers. As to be expected, the numbers of citations, cited authors, and cited publications are higher for conference series with more publications. Interesting is however, that the average reuse of cited references is roughly the same (between 1.60 and 1.80 for all three series individually and 1.94 overall). The relatively low average is an indication that most of the referenced publications are referenced only once. The fraction of 20–22 % self-citations is comparable to or lower than previous reports: Aksnes [2003] reported a share of 21 % within a corpus of publications with at least one Norwegian author and a share of 22 % for only the

Table 3.2: The history of the three FCA-minded conference series in numbers.

	ICCS	ICFCA	CLA	total
publications at the conferences				
editions	19	9	7	35
publications	567	208	179	954
avg. publications per edition	29.84	23.11	25.57	27.26
authors	542	218	269	872
avg. publications per author	2.04	1.94	1.62	2.25
citations <i>by</i> the conferences' papers				
citations	10,131	4,328	2,662	17,121
cited authors	5,871	2,655	2,027	8,513
cited publications	6,079	2,406	1,668	8,813
avg. citations per cited publication	1.67	1.80	1.60	1.94
self-citations	2,255	965	529	3,749
self-citations (share in %)	22	22	20	21
citations <i>to</i> the conferences' papers				
citations	3,569	1,674	209	5,452
citations per publication	6.29	8.05	1.17	5.71
citing publications	2,005	1,281	184	3,035
cited publications	423	136	52	611
cited publications (share in %)	75	65	29	64

mathematical publications in that corpus. Thijs and Glänzel [2006] reported 38% for mathematical publications in a corpus of publications from European universities. Such self-citation statistics are, however, only roughly comparable: Particularly, we count the self-citation rate synchronously (i.e., among the references of publications) whereas both Aksnes [2003] and Thijs and Glänzel [2006] conducted diachronous studies, counting self-citations among the citations a publication received. The third part of the Table shows statistics on citations *to* publications of one of the three conference series.<sup>9</sup> ICCS has received many more citations than ICFCA and the latter more than CLA, which can partly be explained by the age difference and thus the publication volumes of the conferences. However, we also observe that, compared to publications of CLA, a larger share of ICCS and ICFCA publications have been cited at least once; and furthermore, that both these conferences have much higher citation

<sup>9</sup>The numbers of received citations are higher than those reported by Doerfel et al. [2012b] (the paper in which most of the results in this chapter were published). The differences result from retrieving these citations again in 2014, thus including a larger time-window in which they could have occurred. On the other hand however, the matching criteria to compare queried and found publication are now more strict than before: Among others, we verified the venue reported in Microsoft Academic Search and manually checked various search results.

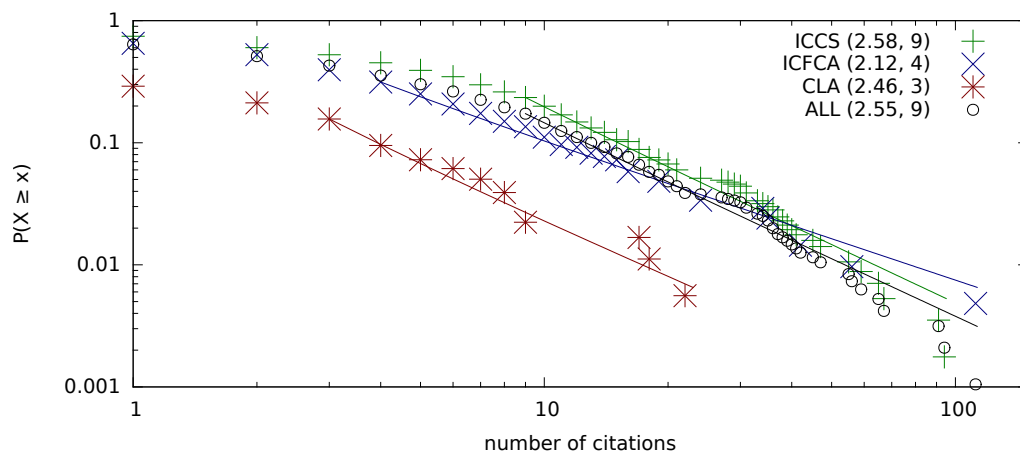


Figure 3.1: Citation frequency distribution of the three conferences and the full corpus. Shown are the cumulative probability distributions together with their power-law fits. The numbers in the key denote the parameters  $\alpha$  and  $x_{min}$  of the power-law fit.

rates per publication (6.29 and 8.05 versus only 1.17 for CLA). This is an indicator for higher impact and better visibility of ICFCA and ICCS.

To investigate the distributions of citations to publications from the three conferences, we use the methods of Clauset et al. [2009], described here in Section 2.1.1, for fitting them to power-law distributions. Power laws are found to be the results of many man-made processes [Clauset et al., 2009], and it has been assumed that citation counts to publications follow power-law distributions (e.g., by Redner [1998] and An et al. [2004]). Figure 3.1 shows the three cumulative frequency distributions of citations to contributions from either conference and from the full corpus (ALL). For each distribution, additionally, the best fit to a power law is depicted. Table 3.3 shows the fits and their goodness-of-fit scores. We see that all four fits are plausible, since their p-values are higher than 0.1.<sup>10</sup>

The weakest descent is measured for ICFCA, where the power-law fit's exponent is  $\alpha = 2.12$ . The other two conferences have slightly higher values. The highest uncertainty for  $\alpha$  is measured for CLA, the conference with the smallest set of observations. Here, also the share of observations that is covered by the fitted power-law model is lowest. Overall, the behavior is dominated by ICCS which is no surprise as it accounts for more than two thirds of the cited publications.

Our results for  $\alpha$  are lower than those found in previous studies that investigated power-law fits in other publication corpora: Albarrán and Ruiz-Castillo [2011] determined exponents  $2.92 \leq \alpha \leq 5.05$  for various scientific disciplines, using a corpus of

<sup>10</sup>Remember that the p-value for a fit, proposed by Clauset et al. [2009], counts a share of synthetic datasets where the goodness-of-fit score is worse than that for the empirical dataset. Thus high p-values are an indication for plausibility of a fit. See Section 2.1.1.

Table 3.3: Power-Law fits to the distributions of citations to the three FCA-minded conferences. The table shows the estimated parameters  $\alpha$  and  $x_{min}$  together with their uncertainties (sd.), the goodness-of-fit (gof), which is the Kolmogorov-Smirnov-statistic between the empirical data and the fit, together with its p-value, and the share of observations that is actually fitted to the power law (cf. Section 2.1.1).

	$\alpha \pm \text{sd.}$	$x_{min} \pm \text{sd.}$	gof	p-value	share in %
ICCS	$2.58 \pm 0.26$	$9 \pm 3.35$	0.05	0.35	23.46
ICFCA	$2.12 \pm 0.29$	$4 \pm 1.99$	0.06	0.48	31.40
CLA	$2.46 \pm 0.41$	$3 \pm 0.88$	0.07	0.63	15.64
ALL	$2.55 \pm 0.26$	$9 \pm 3.57$	0.05	0.18	17.31

citations to journal articles from the Web of Science, and found mathematics and computer science – the two disciplines in their study that are the most related to FCA – to be in the lower part of the interval ( $\alpha = 2.83$  and  $\alpha = 2.92$ , respectively). Similarly, Brzezinski [2015] determined exponents  $2.78 \leq \alpha \leq 4.69$  and particularly 2.78 for mathematics and 3.11 for computer science, using a corpus of journal articles from Scopus. In both studies, the optimal values for  $x_{min}$  were much higher than those measured here: 18 and higher for mathematics and computer science and between 18 and 152 [Albarrán and Ruiz-Castillo, 2011] or between 12 and 148 [Brzezinski, 2015] in general. Possible explanations for these differences might lie in the different corpora (both mentioned studies used journal publications and their recorded citations in expert-controlled catalogs), different citation windows or indeed different citation behavior. The latter cannot be ruled out considering the already broad spectrum of fits found for the different disciplines in both studies.

A comparison with four other candidate distributions (exponential, lognormal, and Weibull distribution, as well as a power law with exponential cut-off) showed that power-law fits are better than those to exponential distributions. The difference is significant (with p-value below 0.1) only for the full corpus ( $p = 0.09$ ). The other three candidates yield better fits than power law, especially the power law with exponential cut-off is a significantly better fit for ICCS, ICFCA, and for the full corpus. By and large, these results are evidence for heavy-tailed distributions in all cases. These results are consistent with Brzezinski [2015] who also found other heavy tailed distributions and particularly power laws with exponential cut-off to be better fits than plain power-law distributions. Regarding the comparison of the three conference series, the distributions confirm that publications of ICCS and ICFCA are more likely to be cited than those of CLA.



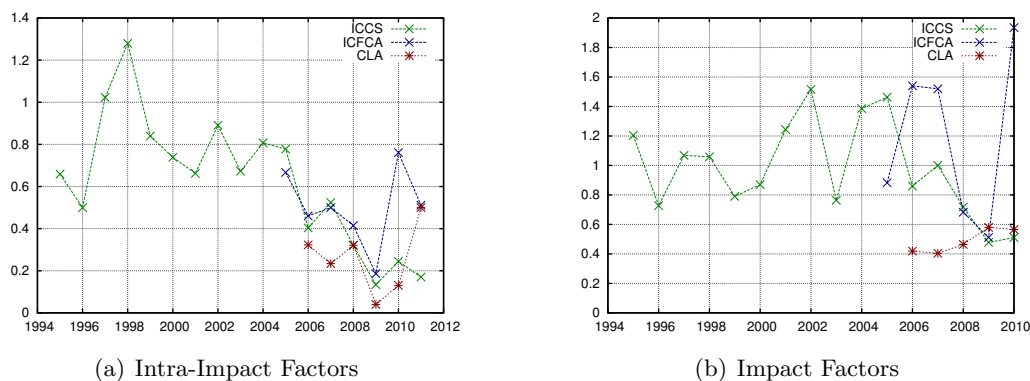


Figure 3.2: The (two-year) impact factors for each conference series over the years.

### Impact of Individual Editions of Each Conference

We take a look at the impact factors (cf. Section 2.2.1) of the conference series. Additionally, we also consider what we call *intra-impact factors*. These are the impact factors computed only from the citations between publications in the corpus. Thus, the intra-impact factor is an estimate for a conference’s impact on the community itself, whereas the regular impact factor is an indicator for its scientific impact in general. Figure 3.2 shows both impact factors for each conference series in every<sup>11</sup> year.<sup>12</sup> In both cases, the impact factor fluctuates heavily, however, some trends are clearly visible. The intra-impact factors of ICCS decline heavily following the start of ICFCA (and one year later CLA) conferences. This corresponds to the intuitive assumption that the new platforms scatter the focus of the FCA community over three series instead of only one. Comparing impact factors over time is generally difficult since the numbers of publications and citations change. Therefore, we compare impact only per year. Almost in every year, the ICCS and ICFCA conferences score higher impact factors (both types) than the CLA conference although the later editions have scores comparable or better than ICCS (and ICFCA in 2009). Especially inside the community, the ICFCA conferences have had a higher impact than CLA.

In Figure 3.3, we directly show which edition of which conference had an impact on the community in which year. For each pair of years  $(x, y)$  there is a pie chart that shows the citations from any conference in year  $x$  to papers from either of the

<sup>11</sup>The impact factors in Figure 3.2(b) stop in 2010 since it was found that in 2011, the number of publications indexed in Microsoft Academic Search dropped (see [Orduna-Malea et al., 2014] or our discussion in Section 6.4.3)

<sup>12</sup>By definition, the intra-impact factor must always be lower than the impact factor, as long as both are measured using the same set of citing publications. In our study this was, however, not the case since the two computations use different sources to count the citations: Figure 3.2(a) uses the references from publications in the corpus and thus reflects the accurate values. Figure 3.2(b), however, uses data retrieved from Microsoft Academic Search. Since the search engine does not cover all publications, its citation counts underestimate the true numbers of citations.

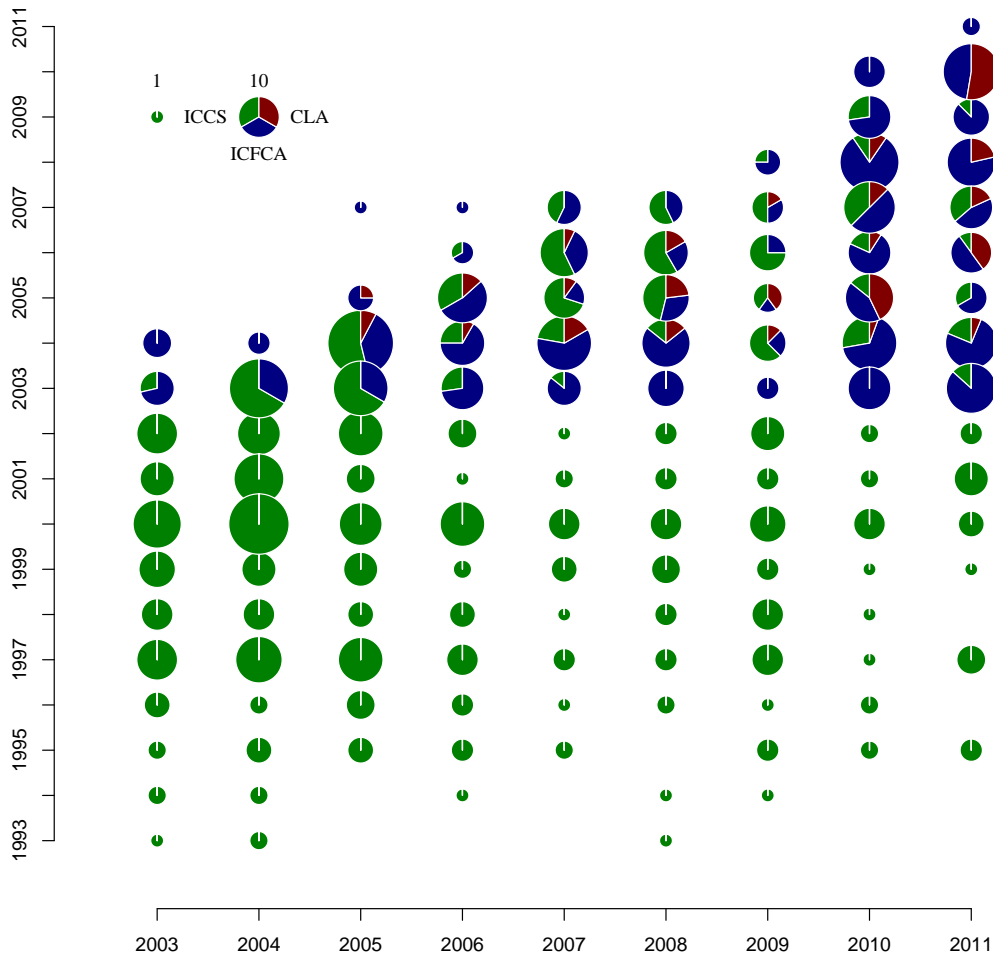


Figure 3.3: Citations from and to papers of ICCS, ICFCA, and CLA over the years. A pie chart positioned at  $(x, y)$  shows the citations from papers of any of the three conferences ICCS, ICFCA, or CLA in year  $x$  to papers of these conferences in year  $y$ . The size of the charts is proportional to the number of citations. The sections of each chart show the shares of each conference among the cited papers.

conferences in year  $y$ . While the size of each pie chart is proportional to the number of such citations, the shares illustrate the impact of each conference in year  $y$  to the community in year  $x$ .<sup>13</sup> As an example, we see that among the references in papers

<sup>13</sup>One would expect that citations can only refer to publications from the same or from past years. However, there are a few exceptions: Twice (in 2005 and in 2006), an unpublished manuscript was cited, that appeared later in 2007 at ICFCA as [Ganter, 2007], and thus in our corpus the year was set to 2007 for these references. In 2003, ICFCA had no published proceedings but instead, selected papers were published in [Ganter et al., 2005] (see Section 3.4.1). Thus, it was possible to include references to literature from 2004.

from 2011 to publications from any of the three conferences in 2003, the major share are references to publications from ICFCA 2003 while the others reference publications from ICCS 2003. There are no references to CLA 2003 since the earliest CLA took place in 2004. The diagram starts on the  $x$ -axis with the year 2003 since before, all citations were from ICCS to ICCS as it was the only conference series at that time. Thus obviously, the citations to publications of years before 2003 are all ICCS only (the lower part of the diagram).

We can observe that among the recent citations of each year, ICFCA quickly gained influence. In 2006, most of the citations to any of the three conferences in years 2003 and later referenced publications from ICFCA. Especially in 2010 and 2011, ICFCA accounted for the largest part of references to publications younger than 2003. Beginning in 2005, CLA also started to influence publications, however, not as pronounced as ICFCA. The influence of ICCS is visible in all years since publications from the earlier years of ICCS have been cited. Over time however, ICFCA started to dominate ICCS and finally, the 2010 edition of ICCS has not been cited by any paper of the three conferences in 2011.

### Publication Habits

Since the three series have been running in parallel over several years, there has been opportunity for authors both to publish several times at the same series and also to publish at more than one series. The most prolific authors of each series (and in total) are listed in Table 3.4.

To gain insights into the publication habits, we now apply FCA to data from our corpus. We consider the many-valued context  $\mathbb{K}_{\text{conf}}^f$  (see Section 3.4.2) to distinguish one-time participants from frequent visitors of the conference series. Through conceptual scaling this context is transformed into the single-valued context

$$\mathbb{K}_{\text{conf}} = (A, \{CLA, ICCS, ICFCA, 3 \times CLA, 3 \times ICCS, 3 \times ICFCA\}, I_{\text{conf}}),$$

where authors have the attributes ICCS, ICFCA, or CLA if they published there at least once. Authors have one of the other three attributes if they published at the respective conference at least three times. More formally, for the attribute ICCS of  $\mathbb{K}_{\text{conf}}^f$ , we use the scale context  $\mathbb{S}_{ICCS} := (\mathbb{N}, \{ICCS, 3 \times ICCS\}, J_{ICCS})$ , where for  $n \in \mathbb{N}$ :

$$(n, ICCS) \in J_{ICCS} \iff (n \geq 1) \quad \text{and} \quad (n, 3 \times ICCS) \in J_{ICCS} \iff (n \geq 3).$$

The scales for ICFCA and CLA are defined analogously, and applying all three scales to  $\mathbb{K}_{\text{conf}}^f$ , we yield  $\mathbb{K}_{\text{conf}}$ . The threshold of three was selected since publishing three times at the same conference series indicates a certain commitment to that series. On the other hand, we did not set a higher value since especially CLA and ICFCA are young conferences, with only seven and nine editions, respectively. The line diagram of the context's concept lattice is depicted in Figure 3.4, where the values below each concept

Table 3.4: The top five contributing authors of each conference with their number of publications. In case of a tie, all authors with the same number of publications are listed.

ICCS		ICFCA	
R. Wille	24	R. Wille	14
G.W. Mineau	19	P. Eklund	11
J.F. Sowa	14	P. Valtchev	10
S.O. Kuznetsov	13	B. Ganter	10
M. Keeler	13	S.O. Kuznetsov	8
		S. Ferré	8
		L. Nourine	8

CLA		total	
S. Ben Yahia	13	R. Wille	42
R. Bělohlávek	11	S.O. Kuznetsov	27
A. Napoli	10	P. Eklund	26
E. Mephu Nguifo	8	B. Ganter	24
V. Vychodil	7	P. Valtchev	20
M. Huchard	7	G.W. Mineau	20
J. Outrata	7		

count the number of authors in the concept extent (support values). Exemplarily, the top contributing authors (Table 3.4) are annotated at their object concepts.

To interpret the lattice, one has to keep in mind that ICCS runs more than twice as long as the other two conference series, naturally resulting in higher author participation: 542 authors versus 218 at ICFCA and 269 at CLA. Yet, the share of authors who returned frequently to the same series is highest for ICFCA: 46 out of 218 authors, approximately 21.10%. Despite the longer run time of ICCS, the rate here is 17.53% and thus lower than that of ICFCA, albeit higher than that of CLA (11.90%).

Most authors published at only one of the conference series: of the 872 authors, only 127 (14.56%) published at least at two series.<sup>14</sup> Only 30 authors (3.44%) published at all three conference series. The overlap between two conference series is smallest for CLA and ICCS (42 authors) and largest for ICFCA and CLA (78). This fits well to the

---

<sup>14</sup>The number 127 is not directly annotated in the line diagram as it does not belong to one specific concept. However, it is easily computed using the sieve formula as:  $127 = 67 + 42 + 78 - 2 \cdot 30$ , 30 being the number of authors who published at all three conferences, 67 the number of authors who published at both ICCS and CLA, and 78 and 42 being the according numbers for the other two combinations.

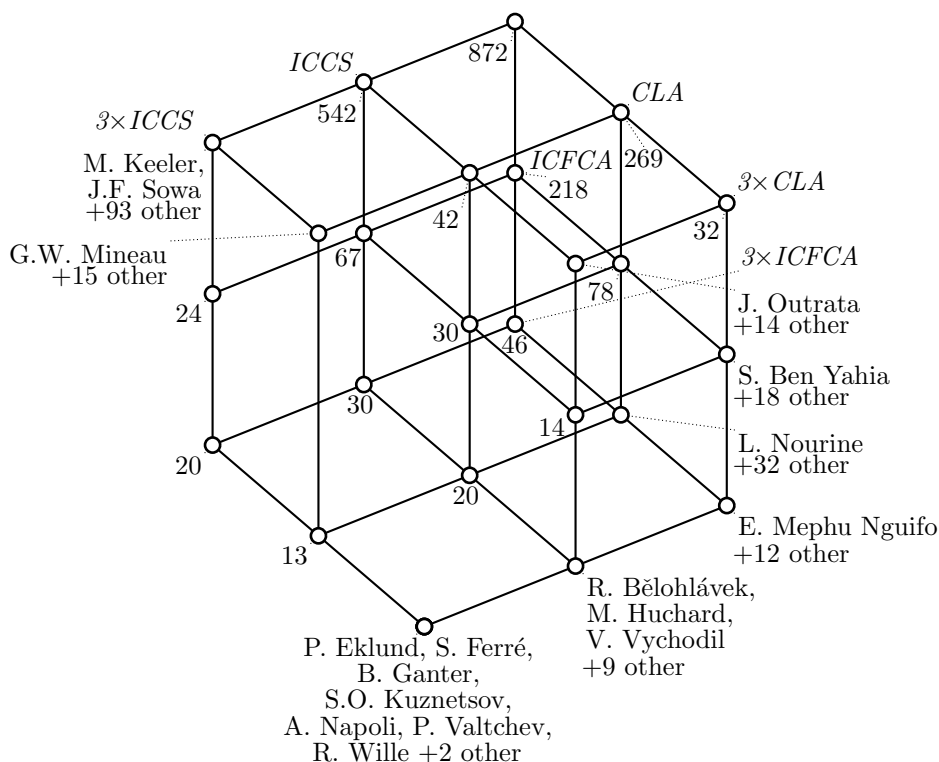


Figure 3.4: The concept lattice for the author-conference context  $\mathbb{K}_{\text{conf}}$ , annotated with support values and the top contributing authors listed in Table 3.4.

fact that while ICFCA and CLA are strictly FCA-minded, ICCS is also a conference for the conceptual graphs community.

Next, we compare, how many of the authors of one series published also at one of the other two series. We observe that this share is highest for ICFCA (115 out of 218 authors, 52.75%), followed by CLA (90 out of 269, 33.46%), and lowest for ICCS (79 out of 542, 14.58%). The same shares computed among authors who frequently visited a particular series also reflect this order: 43 out of 46 frequent ICFCA contributors (93.48%) published also at ICCS or CLA, 20 out of 32 frequent CLA contributors (62.50%) published at ICCS or ICFCA, and 27 out of 95 frequent ICCS contributors (28.42%) published also at ICFCA or CLA. One possible explanation for the relatively low values of ICCS, compared to ICFCA and CLA, is that ICCS, in contrast to the other two series, is not focused on FCA only. Thus, while some authors who published FCA results at ICCS began publishing at ICFCA and CLA when these series started, those authors whose research is dedicated to other areas (mainly conceptual graphs) had no reason to switch to ICFCA or CLA. We also note that the shares among frequent contributors are much higher than those among all authors of a conference

series, which indicates that researchers who are more than casually invested in one of the conferences tend use the other two platforms as well to publish their results.

The Duquenne-Guigues basis of implications (cf. Section 3.2) for  $\mathbb{K}_{\text{conf}}$  contains – aside from the trivial rules resulting from the choice of scales – only two members:

- $3 \times \text{ICCS}$  and  $3 \times \text{CLA} \implies 3 \times \text{ICFCA}$
- $3 \times \text{ICCS}$  and  $\text{ICFCA}$  and  $\text{CLA} \implies 3 \times \text{ICFCA}$  .

The first rule states that any author who frequently published at both ICCS and CLA also frequently published at ICFCA. Similar rules do not hold for the other combinations of conferences. The second rule states that authors who are strongly invested in ICCS, and who published also at both ICFCA and CLA are also frequent contributors to ICFCA.

Several association rules<sup>15</sup> with high confidence, mined from  $\mathbb{K}_{\text{conf}}$ , further confirm the bonds between the three conferences; particularly for frequent participants. In the following, we list the rules with a confidence greater or equal to 80 % (each with its absolute support, counting the authors to whom the rule’s premise applies, and confidence):

- $3 \times \text{CLA}$  and  $\text{ICCS} \implies \text{ICFCA}$  (15/93 %)
- $3 \times \text{CLA}$  and  $3 \times \text{ICFCA} \implies \text{ICCS}$  (13/92 %)
- $3 \times \text{CLA}$  and  $\text{ICCS}$  and  $\text{ICFCA} \implies 3 \times \text{ICFCA}$  (14/86 %)
- $3 \times \text{ICCS}$  and  $\text{ICFCA} \implies 3 \times \text{ICFCA}$  (24/83 %)
- $3 \times \text{ICCS}$  and  $\text{CLA} \implies 3 \times \text{ICFCA}$  (16/81 %).

Roughly speaking, these rules express the fact that many authors who frequently published a paper at ICCS or CLA also (frequently) published a paper at ICFCA.

### 3.5.2 Authors

In this section, we analyze the three conferences on the level of authors. We address Research Question RQ2 – regarding collaboration and influences for (groups of) authors – by investigating co-authorships and citations. Afterwards, we focus on the identification of particularly exceptional authors and their roles, tackling Research Question RQ3. Next to data-mining techniques like clustering or graph-based centrality and role metrics, we use FCA to visualize influences and to determine frequent collaborations.

---

<sup>15</sup>The problem of mining association rules from a database of transactions has been introduced by Agrawal et al. [1993]. It is a relaxation of the implication mining problem: in contrast to implications, association rules must not hold for all instances. Given a set of items and a set of transactions, in which these items occur together, the task is to determine rules that hold with high confidence and are applicable to a large number of transactions (support). Efficient computation algorithms using FCA have been presented by Pasquier et al. [1999] and Stumme et al. [2002].

### Frequent Collaborators

The most frequent collaborators can be read-off from an iceberg lattice of the publication-author-context  $\mathbb{K}_{pa}$ . Given a support threshold (in this case, a minimum number of publications), the iceberg lattice (see Section 3.2 or for more details [Stumme et al., 2002]) of  $\mathbb{K}_{pa}$  contains those concepts where the cardinality of the extent meets or exceeds that threshold. Setting for instance the minimum support (i.e., the minimum number of publications) to six, the following ten pairs constitute the only non-singleton intents of the iceberg lattice (given with their absolute support):<sup>16</sup>

- R. Bělohávek/V. Vychodil (10)
- S. Ferré/O. Ridoux (9)
- J. Ducrou/P. Eklund (8)
- M.R. Hacene/P. Valtchev (8)
- P. Øhrstrøm/H. Schärfe (8)
- R. Godin/P. Valtchev (7)
- E. Mephu Nguifo/S. Ben Yahia (7)
- M. Ducassé/S. Ferré (6)
- B. Ganter/S.O. Kuznetsov (6)
- T. Hamrouni/S. Ben Yahia (6).

The fact that only pairs show up means that there were no teams of three or more authors who published more than six papers together. Using a lower minimum support threshold of 4 yields another 12 concepts with 5 publications and 8 concepts with 4 publications in the extent. Among them are three concepts with intents containing more than just two authors:

- P. Cellier/M. Ducassé/S. Ferré (5)
- T. Hamrouni/E. Mephu Nguifo/S. Ben Yahia (5)
- M.R. Hacene/M. Huchard/P. Valtchev (4).

### Co-Authorship Clusters

The co-author graph  $\mathfrak{G}_{coa}$  reveals interesting patterns of collaboration within and between the FCA and CG (Conceptual Graphs) communities. The map in Figure 3.5 shows a graph clustering (or community allocation, see Section 2.1.2), created using GMap [Gansner et al., 2009]. Connected components that contain less than four authors or that are based on less than four papers have been omitted for the sake of legibility. Thus, we see 29 clusters, that contain between 4 and 65 authors. The width of an edge between two co-authors reflects the number of publications they have written together at any of the three conferences; similarly, the size of the author names depicts the number of published papers. The coloring of author names is related to certain roles they play and will be discussed at the end of this section. The alignment

<sup>16</sup>We omit the line diagram of the iceberg lattice since it is structurally just an anti-chain.

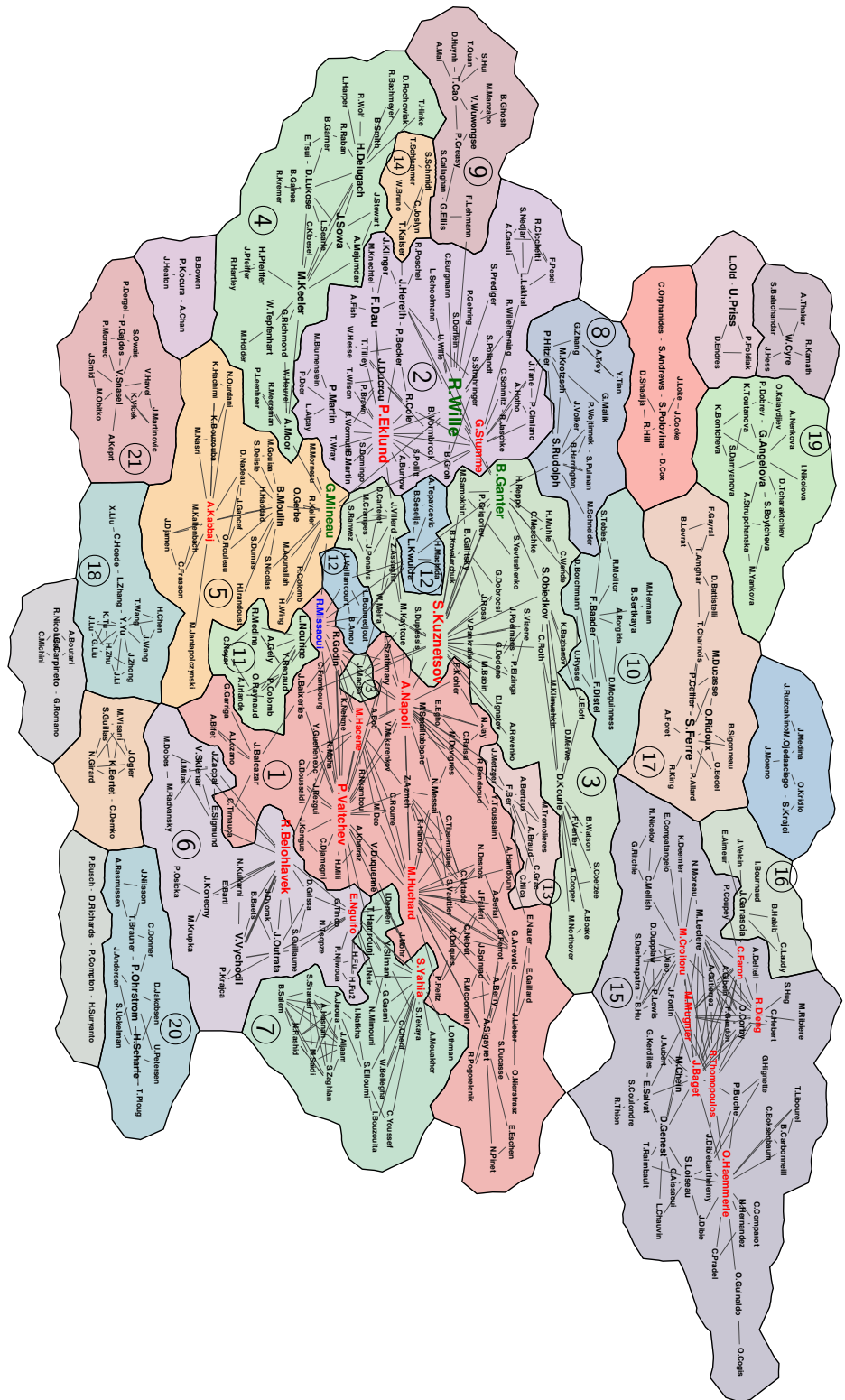


Figure 3.5: A map of the co-author graph. Isolated ‘islands’ (clusters) with less than four publications or less than four authors have been omitted. (Clusters 3 and 18 have been split in the map and therefore they are annotated twice.)



metrics of Scripps et al. [2007a] – cf. Definition 2.5 in this thesis – are relatively high:  $p = 0.96$  and  $q = 0.94$ . This means that the community allocation fits well to the graph structure of  $\mathfrak{G}_{\text{coa}}$  but also that there are some edges that run between communities ( $p < 1$ ), indicating that there are authors who bridge communities through their collaboration with others.

The giant connected component (GCC) of the graph is divided into 14 clusters – numbered 1 through 14 in Figure 3.5 – and contains 321 of the 482 authors shown on the map. The second largest component contains only two clusters: 15 and 16. The remaining five large clusters with more than ten members (17 through 21) are each smaller, single components of the graph, that are not connected to the outside. Judging from the conferences at which the authors of a cluster have published, as well as by looking at the papers in each co-authorship cluster, we can roughly group them into rather FCA-related clusters: 1–3, 6, 7, 8, 10–14, 17, and 21, and CG-related clusters: 4, 5, 9, 15, 16, and 18–20. The two connected components containing several clusters are thus one CG-related component (Clusters 15 and 16) and the GCC, which is composed of both rather FCA-minded and CG-related clusters of authors. To determine more fine grained topics for each cluster, one would have to investigate the actual content of the papers (e.g., applying text mining to identify suitable words that distinguish one cluster from others). However, that would require access to full-texts of every publication, which is beyond the intended goal of this analysis.

## Influences

In this section, we investigate the influences of author communities. Traces of influence can be obtained from the references of publications since citing another publication can be interpreted as evidence for impact of that publication and thus also for impact of its authors. These citation counts can be read of the many-valued contexts  $\mathbb{K}_{\text{cp}}^f = (C_k, P, \mathbb{N}, I_{\text{cp}}^f)$  and  $\mathbb{K}_{\text{ca}}^f = (C_k, A, \mathbb{N}, I_{\text{ca}}^f)$ , introduced in Section 3.4.2.

For legibility, we set the parameter  $k$  to  $k = 8$ , thus using the set  $C_8$  of the eight largest clusters as objects.<sup>17</sup> These are Clusters 1–7 and 15:

- 1 (P. Valtchev, A. Napoli, A.M.R. Hacene, ...),
- 2 (R. Wille, P. Eklund, F. Dau, ...),
- 3 (S.O. Kuznetsov, B. Ganter, S. Obiedkov, ...),
- 4 (J.F. Sowa, H.S. Delugach, M. Keeler, ...),
- 5 (G.W. Mineau, B. Moulin, A. Kabbaj, ...),
- 6 (R. Bělohávek, V. Vychodil, E. Mephu Nguifo, ...),

<sup>17</sup>Each cluster in  $C_8$  contains more than 24 authors while the other clusters contain each at most 14 authors.

- 7 (S. Ben Yahia, T. Hamrouni, Y. Slimani, ...) and
- 15 (J.-F. Baget, O. Haemmerlé, M.-L. Mugnier, ...).

Furthermore, to focus the analysis on only the most influential publications and authors, we reduce the number of attributes by choosing suitable subsets from  $P$  and  $A$ . To select such subsets, two choices suggest themselves: one can either use the most relevant entities (cited publications or authors) globally, or one can choose them per cluster and then merge them together into one attribute set.<sup>18</sup> We illustrate the first choice for the case of cited publications and the second one for cited authors: With the set  $P_{20}$ , we select the 20 most often cited publications of the corpus. In contrast to that, the set  $A_{C_5}$  contains for each of the eight clusters its top five most often referenced authors, where references are counted over all publications with at least one author from the cluster.

Eventually, we derive single-valued contexts using conceptual scaling: For each publication  $p \in P_{20}$ , the scale  $\mathbb{S}_p$  is a context  $(\mathbb{N}, \{p\}, I_p)$ , where  $(n, p) \in I_p \iff n \geq \tau$  for some fixed threshold  $\tau$ . Thus in the resulting single-valued context  $\mathbb{K}_{cp} = (C_8, P_{20}, I_{cp})$ , a cluster  $c \in C_8$  is related to a publication  $p \in P_{20}$  if the authors from the cluster cited  $p$  at least  $\tau$  times; more formally:  $(c, p) \in I_{cp} \iff p(c) \geq \tau$ . Analogously, the context  $\mathbb{K}_{ca}$  is constructed.

Reasonable thresholds  $\tau$  have to be fixed to decide when a cluster is set in relation with a cited publication in  $\mathbb{K}_{cp}$  or respectively with a cited author in  $\mathbb{K}_{ca}$ . There is no rule to select this parameter and it will depend on the user of this method to make a suitable choice or to experiment with various parameters, thus analyzing different levels of influence. In this analysis, we set the threshold  $\tau$  to three for publications in  $\mathbb{K}_{cp}$  and to five for authors in  $\mathbb{K}_{ca}$ , respectively; meaning that a cluster  $c$  is set in relation with a publication  $p$  (an author  $a$ ), if  $p$  ( $a$ ) is cited by at least three (five) papers from  $c$ . The rationale behind choosing these different thresholds is that while per publication, another publication can only be cited once, authors can be cited several times in the same publication (i.e., with several of their publications).

Figures 3.6 and 3.7 show the resulting lattice diagrams. Both lattices seem to reflect the two main schools of thought in the considered conferences: FCA and CG. Each cluster cites one of their cornerstone-publications, [Wille, 1982] for FCA and [Sowa, 1984] for CG, and at least one of their creators R. Wille and J.F. Sowa. Judging from Figure 3.6, Clusters 1, 6, and 7 belong to the FCA community, referencing R. Wille but not J.F. Sowa, and Cluster 15 to the CG community, referencing J.F. Sowa but not R. Wille. Clusters 2, 3, 4, and 5 cite both authors frequently. From Figure 3.7, we see that all clusters but 4, 5, and 15 use the FCA foundations book by Ganter and Wille [1999].

The philosophical foundations of C.S. Peirce are important for Clusters 2 and 4, which are the two clusters around the two founders R. Wille and J.F. Sowa, who constructed their theories with Peirce's work in mind.

---

<sup>18</sup>To yield valid contexts, the relations  $I_{cp}^f$  and  $I_{ca}^f$  must be restricted accordingly.

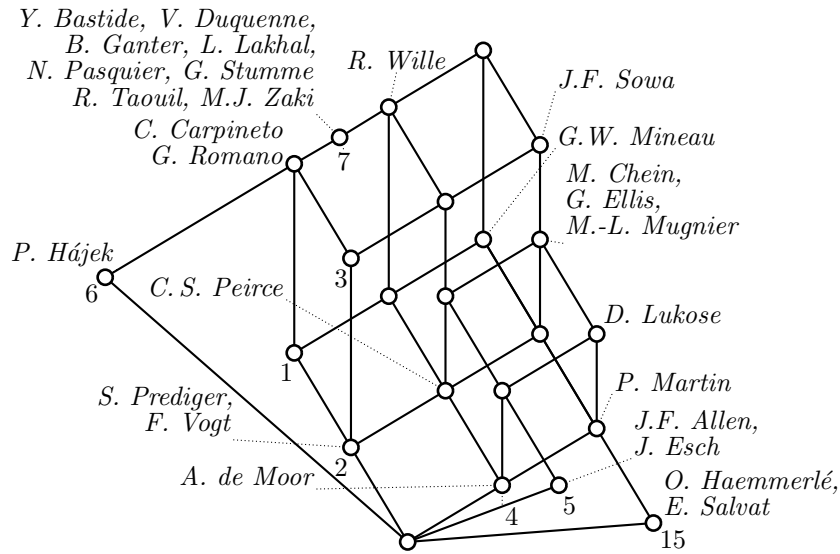


Figure 3.6: The concept lattice of  $\mathfrak{B}(\mathbb{K}_{ca})$ , showing influential authors (as attributes) for the eight largest co-authorship clusters (objects).

In the FCA community, we can see the high impact of the foundations book by Ganter and Wille [1999] (six of the eight author clusters cited it frequently) and of papers on implications [Guigues and Duquenne, 1986] (five clusters) and association rules [Stumme et al., 2002] (four clusters). The topics of the papers further suggest that Clusters 1, 3, 6, and 7 often cite important algorithmic FCA publications [Bordat, 1986, Kuznetsov and Obiedkov, 2002, Ganter, 1984, Carpineto and Romano, 2004, Stumme et al., 2002] – an indication for the strong focus on applicability of FCA notions. Clusters 2, 4, 5, and 15 heavily cite fundamental literature on CG [Sowa, 1984, 2000, Chein and Mugnier, 1992, Mugnier and Chein, 1996]. Cluster 5 and even more so Cluster 2 are also influenced by FCA-minded publications; other than Clusters 5 and 15, both cite Wille [1982].

By and large, we see that Clusters 1, 3, 6, and 7 are FCA-minded, Clusters 4, 5, and 15 are (rather) CG-minded clusters, and Cluster 2 is influenced by both communities. The latter is also confirmed through the relation between Cluster 2 and the frequently cited publication [Wille, 1997], which is a unified approach to both FCA and CG.

### Roles and Key Players

To identify authors in extraordinary positions we consider the two graphs  $\mathfrak{G}_{coa}$  and  $\mathfrak{G}_{cit}$  (cf. Section 3.4). The co-authorship graph  $\mathfrak{G}_{coa}$  is undirected and represents the social network of collaboration. It is therefore suitable to identify roles of individual authors within the network of all authors. The author-citation-graph  $\mathfrak{G}_{cit}$  is directed, and its edges indicate influence: For an author, being pointed to means being cited.

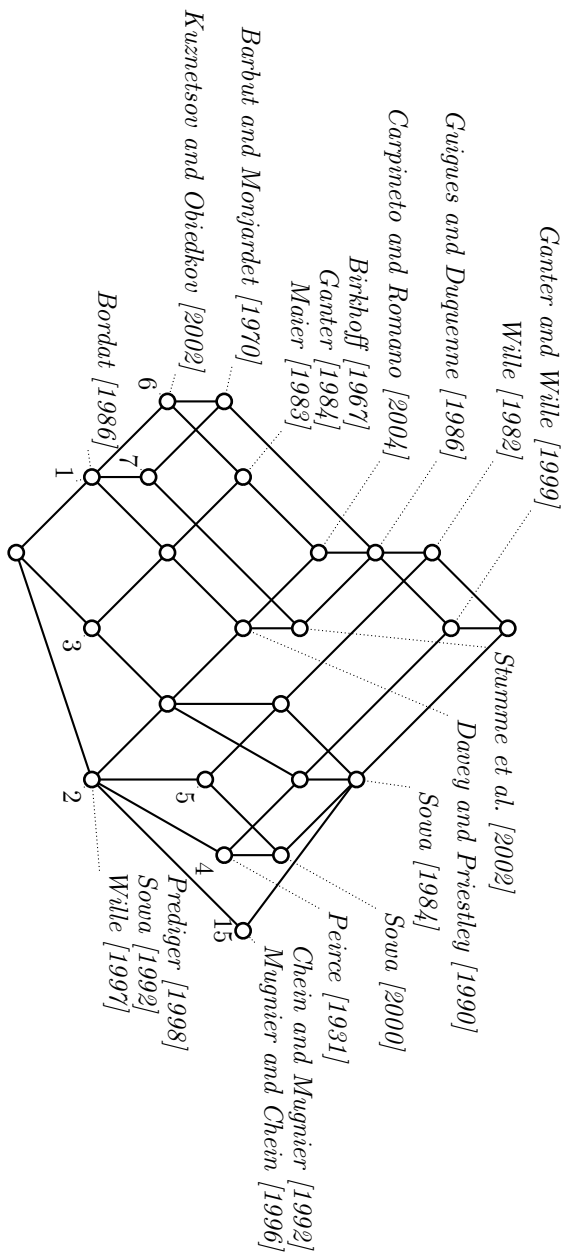


Figure 3.7: The concept lattice of  $\mathfrak{B}(\mathbb{K}_{cp})$ , showing influential publications (attributes) for the eight largest co-authorship clusters (objects).

Table 3.5: Top ten rankings for the graph centrality metrics in-degree, in-strength, *PageRank*, and authority (HITS) in  $\mathfrak{G}_{\text{cit}}$ .

rank	in-degree		in-strength	
1	R. Wille	443	R. Wille	1,877
2	B. Ganter	424	B. Ganter	1,322
3	J.F. Sowa	307	J.F. Sowa	1,033
4	G. Stumme	211	G. Stumme	570
5	R. Godin	156	M.-L. Mugnier	427
6	S.O. Kuznetsov	151	L. Lakhal	412
7	R. Missaoui	134	R. Godin	374
8	G.W. Mineau	128	M. Chein	360
9	L. Lakhal	127	S.O. Kuznetsov	349
10	P. Eklund	124	C. Carpineto	264

rank	<i>PageRank</i>		authority	
1	J.F. Sowa	0.101	R. Wille	0.161
2	R. Wille	0.068	B. Ganter	0.087
3	B. Ganter	0.043	G. Stumme	0.042
4	M.-L. Mugnier	0.021	L. Lakhal	0.031
5	M. Chein	0.020	J.F. Sowa	0.030
6	G. Ellis	0.017	S. Prediger	0.023
7	G. Stumme	0.014	M.J. Zaki	0.019
8	O. Gerbé	0.014	R. Godin	0.019
9	S. Prediger	0.013	S.O. Kuznetsov	0.018
10	G.W. Mineau	0.011	C. Carpineto	0.017

We use  $\mathfrak{G}_{\text{coa}}$  to identify community-based roles and  $\mathfrak{G}_{\text{cit}}$  to identify the most central (most influential) authors. For the latter purpose, we compute four graph centrality measures, explained in Section 2.1.2: In this setting, the in-degree of an author measures by how many authors they were cited and the in-strength counts how often they were cited. The intuition behind using *PageRank* is that authors are influential (have high *PageRank* scores) if they were frequently cited by other influential authors. Finally, we use the authority measure from the HITS algorithm: In the citation graph, authors are good *hubs* if they reference many authors that have high values as authorities (e.g., authors of survey papers would be good hubs); and they are good authorities when they have been cited frequently by good hubs. Thus, authorities are those authors who had high influence on others.

In Table 3.5 we see the top ten positions for these four rankings. One can observe that the different measures show a strong agreement: The forty positions are occupied by a total of only 17 authors, of which twelve (seven) occur in at least two (three) lists. The authors R. Wille, B. Ganter, J.F. Sowa, and G. Stumme even occur in all four lists. The former three are among the top five in each list, which is not surprising as they are the authors of the often cited foundations literature, like [Wille, 1982, Ganter and Wille, 1999, Sowa, 1984].

The co-authorship graph reflects collaboration rather than influence. Here, we can infer exceptional nodes with respect to their position and structure in the community clustering from above, visualized in Figure 3.5. We use the node roles proposed by Scripps et al. [2007b], which have been recalled here in Section 2.1.2. The role of an author  $a$  depends on two values: the node degree of  $a$ ,  $\deg(a)$ , and a community metric. While the node degree can easily be read of the graph, the community metric usually has to be estimated. However, it can also be observed directly from a given community allocation. For that purpose, we use the community allocation  $c : A \rightarrow C$ , where  $C$  are the clusters from Figure 3.5. The community metric for a node (an author)  $a$  is then computed as the number of communities it is connected to. More formally, it is  $|c[\mathfrak{N}(a)]|$ , where  $\mathfrak{N}(a)$  is the set of  $a$ 's neighbors (collaborators).

The authors' roles are determined through comparison to a threshold  $s$  on the node-degree and to a threshold  $t$  on the community metric (cf. Section 2.1.2). Since we are using non-anonymized data, and since we consider only subsets of each author's overall scientific work (through the restriction on the three conference series), we decided to label only the few top actors in the graph. Therefore, we use very high thresholds  $s = 10$  and  $t = 3$ , and we only classify those authors who exceed either threshold. Authors with degree ten or less and with three or fewer communities are not analyzed. Thus, we determine the community role of an author  $a \in A$  by the following condition:

$$role(a) := \begin{cases} \text{Ambassador} & \text{if } \deg(a) > s, |c[\mathfrak{N}(a)]| > t \\ \text{Big Fish} & \text{if } \deg(a) > s, |c[\mathfrak{N}(a)]| \leq t \\ \text{Bridge} & \text{if } \deg(a) \leq s, |c[\mathfrak{N}(a)]| > t \\ \text{unclassified} & \text{if } \deg(a) \leq s, |c[\mathfrak{N}(a)]| \leq t. \end{cases}$$

The community roles of authors are indicated through colors in Figure 3.5 and listed in Table 3.6. In total, 21 authors have been assigned one of the three roles. Nine of these authors, including the three ambassadors and the only bridge, occurred also in Table 3.5 among the most central authors in the author-citation graph.

## 3.6 Conclusion

In this chapter, we have analyzed the citation and collaboration behavior of authors of the three FCA-related conferences ICCS, ICFCA, and CLA. The application of

Table 3.6: Community roles in the co-authorship graph  $\mathfrak{G}_{\text{coa}}$  of the FCA conferences, based on the community allocation from Figure 3.5.

author $a$	deg $a$	$ c[\mathfrak{N}(a)] $
<i>role(a) = Ambassador</i>		
R. Wille	17	4
B. Ganter	12	5
G.W. Mineau	12	4
<i>role(a) = Big Fish</i>		
P. Valtchev	27	3
A. Napoli	24	3
M. Huchard	24	1
S.O. Kuznetsov	23	2
P. Eklund	19	2
E. Mephu Nguifo	16	3
J.-F. Baget	16	1
S. Ben Yahia	15	3
M. Croitoru	15	1
G. Stumme	14	3
R. Bělohávek	14	2
R. Thomopoulos	14	1
M.-L. Mugnier	13	1
O. Haemmerlé	13	1
R. Dieng	13	1
C. Faron	12	2
A. Kabbaj	12	1
M.R. Hacene	12	1
<i>role(a) = Bridge</i>		
R. Missaoui	10	4

classical bibliometrics measures has allowed us to assess and compare the impact of the three conferences series and of individual authors. Through the introduction of FCA to this field, we could go beyond mere quantitative measurements of authors or venues. The constructed contexts and their lattices have been used to identify and visualize interesting patterns of influences (of authors and publications for the community) and conference participation.

Concerning our first research question (RQ1), regarding relations between the different conferences, we saw that the two series ICCS and ICFCA had a higher impact (in terms of citations) than CLA. Among the three conferences, ICFCA assumes a central position: Many contributors of ICFCA contributed also to the other two

conference series and those who frequently published at the other two series also published at ICFCA. To answer Question RQ2, we have not only identified frequent collaborators, but also visualized which authors and which publications have been influential to particular parts of the FCA community. For that purpose, we have constructed suitable formal contexts and computed the line diagrams of their concept lattices. We saw the strong influence of the founders of formal concept analysis and conceptual graphs as well as the grouping of subcommunities into either field. Finally, regarding our last research question (RQ3) on key players and roles, we have enumerated several ways of identifying particularly central authors, and we used a community-based role assignment to group especially remarkable authors by the positions they assume in the community and its subfields. These roles highlight authors who are particularly well connected in a specific subcommunity and those who connect several subcommunities.

Overall, we have created an overview of the FCA community, that can be of use for newcomers in the field, as well as to those who organize the here considered conferences and thus a major part of the FCA community. Newcomers can, for example, get an overview of the various subcommunities and they can choose already established authors to approach for possible collaborations, among others based on the authors' roles or their membership in a particular subcommunity. Conference organizers can estimate the relation of their conference in comparison to the other series or to previous editions. The results regarding individual authors might, for instance, be useful when discussing candidates for invited talks.

With the analyses and visualizations in this chapter, we have demonstrated means to support researchers using data from the creation phase of the scholarly publication life cycle. Methodology from FCA has been shown to be suitable for visualizing relations between conference series, as well as between subcommunities and influential publications and authors. These analyses can be adopted for other research fields as well, supporting researchers in the respective communities.

### 3.6.1 Future Research

All used methodology – and particularly the FCA constructs – can be applied to literature corpora of other communities. The only prerequisite is the compilation of a suitable dataset. In several analyses, we had to choose parameters, like the number of publications at the same conference series to be counted as frequent contributor in Section 3.5.1 or the thresholds  $\tau$  in the scales for the influence contexts  $\mathbb{K}_{cp}^f$  and  $\mathbb{K}_{ca}^f$  in Section 3.5.2. Here, they were chosen rather intuitively, considering the numbers of publications and of editions per conference. However, testing these methods on further datasets could allow the development of more sophisticated ways to choose these parameters or at least rules of thumb for good settings. Moreover, an interactive implementation would enable analysts to experiment with various settings and thus to create more fine-grained results.



Our analysis of the FCA community is restricted to the three conference series ICCS, ICFCA, and CLA. The picture could be completed by adding further publications from journals and books. Finding such relevant publications and retrieving their metadata and citations is clearly a first step for such future work. A suitable start would be the surveys by Poelmans et al. [2013a,b].

While we have analyzed the two levels of conferences and authors, the level of individual publications has been investigated in [Doerfel et al., 2012b].<sup>19</sup> Further investigation could also consider the topic-level. Making use of the full content of the publications instead of just their metadata would allow further analysis using other methodology like text-mining, and could for example reveal deeper insights into the communities of co-authors or discover topic shifts in the scientific discipline of FCA.

---

<sup>19</sup>This part of the analysis has been omitted here as it was mainly the work of co-author Robert Jäschke.



## Chapter 4

### Analyzing Researchers during a Conference



In this chapter, we turn to the second phase of the publication life cycle: dissemination. Using active RFID technology, we were able to track the social interaction between participants of the conference LWA 2010, a workshop event of four special interest groups. The resulting dataset contains the face-to-face contact network between the conference’s participants. Our analysis of these interactions is structured along three levels: The conference itself, the four workshops of the four respective special interest groups, and the individual participants, grouped by their academic position or their function at the conference (organizer or regular participant). We observe the duration and distribution of contacts and various features of the resulting contact network between participants. Moreover, we present an analysis of communities that are induced by the grouping into the four special interest groups, and we discuss their relation to communities of participants that can be mined from the observed contacts using respective community detection algorithms. Furthermore, using community information, we examine different roles that are assumed by the individual participants.

We expect our analyses to be helpful for conference participants to better understand their research community and their own position therein, as well as for the organizers of a conference, who want to create an event that benefits all participants as much as possible. The study we present in this chapter is an adaptation of our work in Atzmueller et al. [2011] and Atzmueller et al. [2012b].

#### 4.1 Introduction

After discussing different conference series by viewing them through their proceedings in the previous chapter, we now study the situation where researchers have chosen a conference and are attending it: From the level of conference series, we zoom in to the level of one event where researchers from different subfields of a research discipline come together. Studying the interactions of researchers at such occasions and what one can learn about the structure of the respective research community, is an open research problem. A better understanding of what researchers actually do at conferences is relevant for those who organize conferences, as well as for the participants themselves. Organizers can adapt their events (e.g., to foster more interaction), and for participants it helps to perceive one’s own status within a community and to compare to others.

The study in this chapter explores the second phase of the scholarly publication life cycle (Section 1.1), dissemination, which probably is the least well studied phase in the cycle. Data of the first and fourth phase of the life cycle (i.e., publication metadata) has been under investigation for quite some time, and data on the usage of publications (third phase) has been investigated since it became available on the web. However, observing interactions of participants of scientific conferences – the events where scholarly publications are presented and thus disseminated –, has rarely been possible so far. Only recently, technology has become available that allows the detection of face-to-face contacts. In this chapter, we describe the utilization of such technology for observing contacts between researchers during a conference for two purposes: (i) to enhance the conference experience for participants by offering the conference guidance system Conferator and (ii) to gain insights into the behavior of the participants.

During a conference, social contacts form an essential part of the participants' experience. Commonly, the term “networking” is used to describe the inherent processes in such interactions. Typically, there are different implicit and explicit communities present at a conference. Explicit communities are induced by some external clustering, for instance by tracks or, like in our use case, by special interest groups, where each member is part of one such track or group. Implicit communities arise from the interactions of the participants and can be detected in the observable contact networks. Furthermore, not all participants of a conference behave alike, they differ in the number of contacts to others, in the number of conversational partners, or in the communities that their partners are members of. These characteristics can be expressed in terms of roles that are attributed to participants based on their position in the interaction networks.

In this chapter, we present an experiment conducted during the computer science conference LWA 2010, a workshop event of four German special interest groups. At this occasion, we deployed the conference guidance system Conferator, which uses active RFID technology to record face-to-face contacts between participants. Conferator provides utility to the participants by offering management tools for talks, as well as a social network component, presenting profiles of other attendees, contact information, and an overview about one's face-to-face contacts during the conference. Thus, Conferator not only assists researchers in planning their personal conference schedule but also helps find interesting dialog partners and keep track of one's conversations during the conference.

**Research Questions.** The system Conferator allowed us to track the participants' interactions for our analysis, which we conduct on three levels to attend to the following research questions:

(RQ1) On the conference level: Which structural properties do the interaction networks of all participants exhibit and how do they compare to results reported previously from another conference (Hypertext 2009, Isella et al. [2011b])?

- (RQ2) On the level of special interest groups: How does community structure given through the partition into the four special interest groups, compare to communities detected in the interaction networks?
- (RQ3) On the level of peer groups: Are there differences in the interaction when comparing researchers grouped by their academic status or their function in the conference (organizers or regular participants)?

Answers to these questions present insights into the social interaction of a conference. They can help participants understand their own position within the community (by comparing their own behavior to that observed in such a study). Such information is also valuable for organizers of similar events as it helps them acknowledge and anticipate the participants' behavior. Moreover, they could also adapt their conference plan based on the results of previous events, for example, by including additional social gatherings, birds-of-a-feather sessions, and so on, to encourage social interaction and the exchange between participants.

**Contributions.** Our contributions in this chapter are three-fold:

1. We introduce the social conference guidance system Conferator, which supports participants in their interactions with their peers as well as in managing and visiting talks.
2. We present an in-depth analysis of the community structures among the participants, identifying communities and comparing them to the community structure induced by four special interest groups.
3. Finally, we determine different roles which participants play in the interaction networks, and we relate them to academic status and the participants' function at the conference.

Using tools like Conferator, organizers can improve the conference experience for their participants by offering recommendations or means to “post-process” the conference and the individual conversations. Analyses like those presented here, can be integrated in such tools and thus allow a just-in-time overview of one's own activities compared to those of the community. Moreover, they could serve as the basis for recommender algorithms (e.g., recommendations of conversational partners).

**Limitations.** The study is limited to one edition of LWA and thus to a relatively small, specific group of people. Thus, the results are only valid for this particular conference and due to the small number of participants, individuals have a high influence on the overall results. However, since conferences in general attract different crowds, depend on the organizers and on the structure (e.g., multiple tracks versus single track versus workshops), it is not to be expected that there are many general results to be found that hold true for all conferences. Rather, the investigations in this chapter contribute

a stepping stone to the analysis of the publication life cycle by demonstrating how a conference can be studied and by providing results as a basis for comparisons with other conferences. Our methods are generally applicable; in fact, meanwhile<sup>1</sup> they have already successfully been applied to analyze a variety of conferences, like the Hypertext 2011 [Macek et al., 2012] or the LWA 2011 and 2012 [Kibanov et al., 2013].

**Structure.** The chapter is structured as follows: In the next section, we discuss the social conference guidance system Conferator and the way it enables capturing human face-to-face interactions. After that, in Section 4.3, we review related work. Section 4.4 summarizes the collected data on the participants and their face-to-face contacts. In Section 4.5, we present and discuss the results of our analysis, including an in-depth analysis and evaluation of real-world conference data. Finally, Section 4.6 concludes the chapter with a summary and directions for future research.

In this chapter, we present results of experiments originally conducted for [Atzmueller et al., 2011] and [Atzmueller et al., 2012b]. However, all findings presented here are the result of new or extended experiments, using an updated dataset. The discussions have been revised and rearranged accordingly. The two earlier publications contain additional analyses (e.g., an approach to characterizing participants' roles using subgroup discovery). These have been excluded from this thesis as they had been contributed mainly by the co-authors of these publications.

## 4.2 Tracking and Supporting Social Interactions at Conferences

The experience of a conference for its participants consists of three phases: Preparation (before the conference), the actual participation at the conference, and activities after the conference. Appropriate talks and sessions need to be chosen before they are attended; at the conference, conversational partners need to be found and approached; and after the event, talks, and discussions need to be memorized. Social contacts during a conference are essential for networking, and are often revisited after a conference, as are the visited talks. All of these steps are supported by the conference guiding system Conferator. We described the system in greater detail in [Atzmueller et al., 2011].

At its core, Conferator consists of two key components: PeerRadar and TalkRadar. PeerRadar provides information about the social contacts, acting context sensitively by considering the location of other conference participants. TalkRadar helps manage conference information, like the conference schedule and talk details. Both components make use of data gathered from RFID tags that are worn by the participants of the conference. A first prototype of Conferator was successfully deployed at LWA 2010 at the University of Kassel in October 2010 – the conference that serves as the use case in this chapter. Conferator has since been developed further in the context of

---

<sup>1</sup>We published most of the work in this chapter earlier in [Atzmueller et al., 2012b] and presented it at the MUSE workshop in 2011.

the project VENUS.<sup>2</sup> It spawned the Ubicon framework<sup>3</sup> for ubiquitous and social networking [Atzmueller et al., 2012a], which has become the basis for Conferator, as well as further applications (e.g., MyGroup<sup>4</sup>, a system to support the daily work of a research group, or WideNoise and AirProbe, two applications that were central in the project EveryAware<sup>5</sup> [Atzmueller et al., 2014]). In the remainder of this section, we describe the main components of Conferator.

### Active RFID Technology

To detect face-to-face contacts between people, Conferator relies on an approach using a new generation of RFID tags [Alani et al., 2009, Cattuto et al., 2010] that has been developed by the SocioPatterns project<sup>6</sup> and the company Bitmanufaktur, and that has been made available open source by now.<sup>7</sup> The main feature of these tags is their ability to detect the proximity of other tags within a range of up to 1.5 meters. Moreover, due to the fact, that the human body blocks RFID signals, not only mere proximity, but actual face-to-face contacts can be detected between persons (who wear such a tag on their front). We consider sightings between two tags as a face-to-face contact if they last longer than 20 seconds. We assume that a contact has ended when the two respective tags did not record another sighting for more than 60 seconds. Every two seconds, the tags transmit packages that can be received by stationary RFID readers, and that are then processed by Conferator to extract the recorded face-to-face contacts. Evaluating the signal strength of the received packages enables another feature of Conferator: room-level localization, which we described in [Scholz et al., 2011].

### TalkRadar

TalkRadar is one of the two interactive components of Conferator: It serves as an interactive, context-aware conference schedule. The main feature of TalkRadar is a browsable list of all talks, which provides a general overview of the conference. The talks are ordered by their start time and can be picked (or unpicked) by a user to arrange a personal schedule. The list of talks can be filtered by track, time, tags, the publications' keywords, or combinations thereof. Furthermore, a personalization filter offers viewing all those talks that the user has picked, or those that another participant has picked (as long as the other user has chosen to share these information with the active user). To help resolve time conflicts during the selection of talks to attend, TalkRadar offers a time line view, plotting talks with their duration against a scalable time axis, visualizing parallel or overlapping talks. Figure 4.1 shows a screenshot of

---

<sup>2</sup><http://www.uni-kassel.de/eecs/iteg/venus/>

<sup>3</sup><http://ubicon.eu/>

<sup>4</sup><http://ubicon.eu/about/mygroup>

<sup>5</sup><http://cs.everyaware.eu/event/overview>

<sup>6</sup><http://www.sociopatterns.org/>

<sup>7</sup><http://get.openbeacon.org/physicalweb.html> (accessed September 25, 2015)

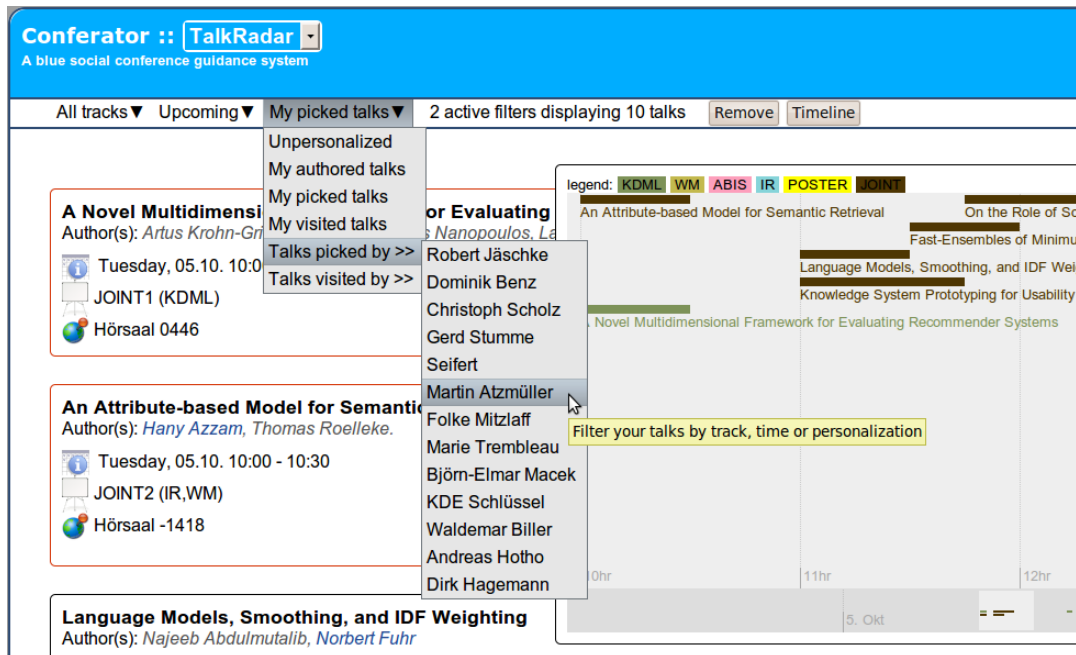


Figure 4.1: A screenshot of Conferator, displaying a the talk list view of TalkRadar. Displayed are talks (two of them highlighted), the time line containing those talks and the filter menu.

the list of talks, the filters, and the time line. During the conference, the list view shows the ongoing and upcoming talks first, where ongoing talks are highlighted. Users attending a particular talk can see who is in the audience with them. The localization component of the RFID setup detects which talks participants are attending. Finally, participants can use TalkRadar to take notes during the talks.

TalkRadar is integrated with the scholarly publication bookmarking system BibSonomy (see Section 2.3.2), allowing a seamless transfer of interesting talks and their respective publications from TalkRadar to one’s own collection there. Thus, talks can be copied, tagged, and retrieved later on. Furthermore, the information and metadata of each publication in the conference proceedings is stored in BibSonomy, and TalkRadar presents it on a descriptive page for each talk. Notes taken during a talk are automatically stored in Conferator, and they are attached to the respective publication posts in BibSonomy as private note.

### PeerRadar

PeerRadar is the second interactive component of Conferator and is at its core a social network. Users of Conferator fill out a form to generate a user profile, including among others their research interests and accounts on other social networks. During the conference, users are provided with an overview of all participants, where they can



browse the profiles of others. What distinguishes PeerRadar from other services is that it can rely on data generated from the RFID tag setup to keep track of actual (offline) social interaction: PeerRadar offers a view that shows a history of all face-to-face contacts a user had during the conference, as well as a list of recent contacts, that is continuously updated. PeerRadar thus helps participants of a conference remember and manage conversations they have had and find those people they want to talk to.

### **Privacy**

In Conferator, privacy is a crucial issue: A variety of user data is collected, including location and contact information. Therefore, appropriate steps for their secure storage and access have been implemented. In Conferator, users manage a list of users they trust. Access to particular pieces of information (e.g., current location, recent contacts, interests) can be restricted to private (nobody else can see), trusted (all users on the trust list can see), or public (every logged-in user can see).

Privacy concerns also pertain the investigations in this chapter. Therefore, in none of the experiments do we analyze individuals, nor do we reveal real names of participants. All results are presented on some level of aggregation, like special interest groups, members of the organizer committee, or participants with the same academic status.

## **4.3 Related Work**

Technology to track what participants of a conference actually do, has only recently become available and thus not too much related literature has been published so far. Therefore, in this section, we review studies that have previously made use of the same RFID framework that we use in this chapter. Furthermore, we briefly discuss other applications that are, like Conferator, dedicated to improve the networking experience for participants of conferences. Since the results of this chapter have originally been published in 2011, a few other studies have already built on our work and continued this line of research. We mention such literature in the future work section at the chapter's end.

### **4.3.1 Conference Applications**

In this chapter, we report results from the deployment of the Conferator prototype, created in 2010. Conferator has since been developed further and has become part of the open source platform Ubicon (see the previous section). An earlier system that applied the same RFID setup was *Live Social Semantics*. It has been deployed at the European Semantic Web Conference 2009 [Alani et al., 2009]. The system collected interaction data and combined it with further information about the participants, gathered from several online sources. Attendees of the conference were supported through means of networking and recommendations of other, potentially interesting participants.

Another system that, like Conferator, supplies conference participants with an online schedule is *Conference Navigator* [Wongchokprasitti et al., 2010, Parra et al., 2012], which allows researchers at a conference to organize their personal conference schedule and offers a lot of interaction features. However, unlike Conferator, it is not connected to the real-life activities of the users during the conference. Conference Navigator has been deployed at various conferences and the collected data has, for example, been used to evaluate talk recommender algorithms [Lee and Brusilovsky, 2014].

Meanwhile, several commercial vendors offer apps for conferences and similar events. Among them are *Whova*,<sup>8</sup> *Pathable*,<sup>9</sup> or *Presdo Match*.<sup>10</sup> These apps incorporate social networking with event guides, including different features, like interactive maps, conference participant directories, messaging, invitations to meetings, user profiles, and integration with other social networks. In contrast to these commercial alternatives, Conferator is an academic project with the primary goal of collecting data about the participants for the purpose of analysis.

### 4.3.2 Analysis of Face-to-Face Contacts

Regarding the tracking and analysis of human interactions, a few approaches have been made using RFID or Bluetooth devices: Hui et al. [2005] used Bluetooth devices, called *iMote*, to measure how often participants of a conference had the opportunity for contacts. Eagle and Pentland [2006] presented an approach for collecting proximity and location information using Bluetooth and GSM signals on mobile phones. They used entropy to measure (ir-)regularity in the mobility patterns of individuals and found pronounced differences between individuals, grouped by their function in the MIT Media Laboratory. They also analyzed proximity patterns between colleagues at work and between friends and found them to be related to office hours.

One of the first applications that used RFID tags to track the position of persons on room basis, was conducted by Meriac et al. [2007] in the Jewish Museum Berlin in 2007. Cattuto et al. [2010] added the aspect of proximity sensing in the SocioPatterns project, where the technology was deployed on a variety of occasions: Isella et al. [2011b] conducted experiments on several contact networks obtained via RFID technology, among others on the Hypertext conference 2009. We compare with their work in this chapter where possible. Barrat et al. [2010] analyzed social dynamics of three conferences and found, for instance, that several statistical properties, like the distribution of contact durations, are very similar across conferences. They also observed that attendees tend to contact people with similar (academic) seniority – measured in terms of the h-index or the number of authored publications – more than others. In their comparison of offline and online social networks (e.g., Facebook and Twitter) they found no correlation between the numbers of people, individuals were connected with. However, participants tended to spent more and longer face-to-face

---

<sup>8</sup><https://whova.com/>

<sup>9</sup><https://pathable.com/>

<sup>10</sup><http://match.presdo.com/>

contacts to participants they were close with in online networks. The RFID setup has also been used in other contexts, like schools [Stehlé et al., 2011] and hospitals [Isella et al., 2011a].

Our work uses the technical basis (RFID tokens with proximity sensing) of the SocioPatterns project, which allows us to generate comparable results. However, in this chapter, we also extend previous experiments by identifying different communities and by discovering roles of participants.

## 4.4 Dataset

Conferator was offered as a service to participants of the conference LWA 2010. For capturing data on social interactions, RFID proximity tags (cf. Section 4.2) of the SocioPatterns project were handed out to the attendees. Out of 100 participants, 71 volunteered to wear an RFID tag and allowed us to track their face-to-face sightings with other participants.<sup>11</sup>

Using the recorded contact data, we generated the undirected networks  $LWA[\geq i]$ , ( $i = 1, \dots, 30$ ). In these networks, an edge  $\{u, v\}$  was created, iff a contact with a duration of at least  $i$  minutes among participants  $u$  and  $v$  had been detected. With the threshold  $i$ , we can filter out short conversations. For instance with  $i \geq 5$  [minutes], we can filter out “small talk” conversations.

LWA 2010 was a joint workshop week of four special interest groups of the German Computer Science Association (GI):

- ABIS, for “Adaptivität und Benutzermodellierung in interaktiven Softwaresystemen” focuses on *adaptivity and user modeling*.
- IR is concerned with *information retrieval*.
- KDML focuses on all aspects of *knowledge discovery and machine learning*.
- WM, for “Wissensmanagement”, considers all aspects of *knowledge management*.

Upon registration for LWA 2010, each participant declared their affiliation to exactly one special interest group. The largest workshop was KDML (37 participants who also took part in the experiment), followed by WM (16), and IR (11). The smallest workshop was ABIS with seven participants, resulting in a total of 71 participants. For all participants, we retrieved their academic position, distinguishing between professors, postdoctoral researchers, PhD students, and students. Participants that had no academic position (e.g., participants from industry) were classified as “other”.

Finally, the organizers of LWA 2010 and the Conferator staff were also among the 71 participants. In the following, we will call all other participants “regular” participants

<sup>11</sup>Six further persons wore RFID tags during the conference, however, they had come as guests particularly for the RFID experiments and did not participate in the conference. Their contacts have therefore been excluded from all analyses.

Table 4.1: Statistics (see Section 2.1.2) for four different contact networks of LWA 2010: Number of (non-isolated) nodes and edges, average degree (avg. deg), average path length (APL), diameter  $d$  (of the largest connected component), density, and clustering coefficient  $C$ , as well as the number of connected components (ignoring isolated nodes)  $\#CC$  and the size of the largest weakly connected component  $|CC|_{\max}$ .

network	$ V $	$ E $	avg. deg	APL	$d$	density	$C$	$\#CC$	$ CC _{\max}$
LWA $[\geq 0]$	71	949	26.73	1.65	3	0.38	0.59	1	71
LWA $[\geq 5]$	67	284	8.48	2.37	5	0.13	0.38	1	67
LWA $[\geq 10]$	60	149	4.97	2.86	6	0.08	0.37	2	56
LWA $[\geq 30]$	18	14	1.56	2.02	4	0.09	0.00	4	8

to distinguish them from attendees who acted both as participants and as organizers of the event. Thus, each participants has three attributes: special interest group, academic status, and organizer or regular participant.

## 4.5 Analysis

In this section, we present our study of the collected interaction data. We analyze the LWA 2010 along three levels: First, the conference with all its participants, then along its tracks and communities, and finally, we investigate roles of individuals in connection to the participants' academic status.

### 4.5.1 Conference

In this part of the analysis, we approach Research Question RQ1 and inspect the conference as a whole. Where possible, we compare the results to a similar experiment at the Hypertext Conference 2009 (HT09), reported by Isella et al. [2011b]. Table 4.1 contains some statistics for LWA $[\geq i]$ ,  $i = 0, 5, 10$ , and 30. The table's first line contains the statistics for the full conference, counting any contacts (longer than 20 seconds, see Section 4.2). A total of 4,992 contacts has been recorded, giving rise to 949 edges between the participants. The graph is relatively dense, forming one giant connected component, with a diameter of only  $d = 3$ . Compared to HT09, LWA 2010 is smaller both in terms of participants and contacts – HT09 had about 100 participants and roughly 10,000 recorded contacts. There, a diameter of  $d = 4$  was measured in a graph of all contacts of one single conference day. The difference might be explained by the shorter duration of the measurement in [Isella et al., 2011b] and the larger number of participants. Another factor could be that most of the participants of LWA are from the same country (Germany) and frequent the LWA series regularly. Thus, many participants already knew each other and could make contact more easily. The average degree of 26.73 at LWA 2010 means that on average, participants had

contact to roughly one third of the other participants. This value is higher than that observed at HT09, which was close to 20, and thus agrees with the assumption of greater familiarity among the participants.

Particularly interesting is the introduction of the minimum threshold  $i$  on the duration of contacts in the networks: By removing all contacts that are shorter than five minutes, four participants become isolated nodes, and overall the number of edges drops to about 30%. The average degree among the non-isolated nodes drops to about one third of the value in  $LWA[\geq 0]$ , and the graph's diameter is two edges longer. A similar effect was observed for HT09 by Isella et al. [2011b]. The effect increases with higher thresholds. Counting only contacts that are longer than ten minutes, the graph is split into two connected components (and eleven isolated nodes). Still, in  $LWA[\geq 5]$  and  $LWA[\geq 10]$ , more than 60 participants are connected to at least one other participant, and we can observe clustering coefficients  $C = 0.38$  and  $C = 0.37$ , respectively, indicating some community structure (see Section 2.1.2 for details). In  $LWA[\geq 30]$ , there are no more three-cliques ( $C = 0$ ), and 53 nodes are isolated.

The influence of the contact length on the various graph features demonstrates that short and long conversations are different indicators. Considering that longer contacts could be interpreted as evidence for more in-depth discussion, it seems plausible to assume that they are more valuable in determining actual community structure. Short contacts, might result from organizational talks or brief small talk conversations. On the other hand, the threshold  $i$  must not be chosen too high, since the graph gets too sparse to be analyzed meaningfully, and it is not unexpected that due to nature of the event, contacts rarely exceed half an hour.

For  $LWA[\geq 0]$ , Figure 4.2 shows the degree and contact length distributions. The latter exhibits characteristics that are comparable to those reported for HT09. Testing fits of different candidate distributions (described in Section 2.1.1) suggests the presence of a power law with exponential cutoff – a distribution proportional to  $x \mapsto x^{-\alpha} e^{-\lambda x}$ , where  $\alpha$  and  $\lambda$  are parameters. The fit is significantly better than a fit to a plain power law and the latter only an insignificantly better fit than the exponential distribution. The degree distributions of the two conferences differ: The distribution at LWA 2010 exhibits three peaks, between five and ten, between 25 and 30, and between 35 and 40. In contrast, the degree distribution at HT09 had only one peak between 15 and 20. One possible factor for this difference is the relatively high number of organizers among the participants. Therefore, we computed a second distribution, this time counting only contacts between regular participants. The resulting distribution has two peaks (between five and ten and between 20 and 30) and is, thus, closer but still different to the one observed at HT09.

By and large, we conclude that, while there are similarities between the social interactions recorded at the two conferences LWA 2010 and HT09, there are also pronounced differences. We can hypothesize that these are due to the different nature of these events. HT09 is an international conference, while LWA is a rather local workshop event, where many of the participants knew each other in advance.

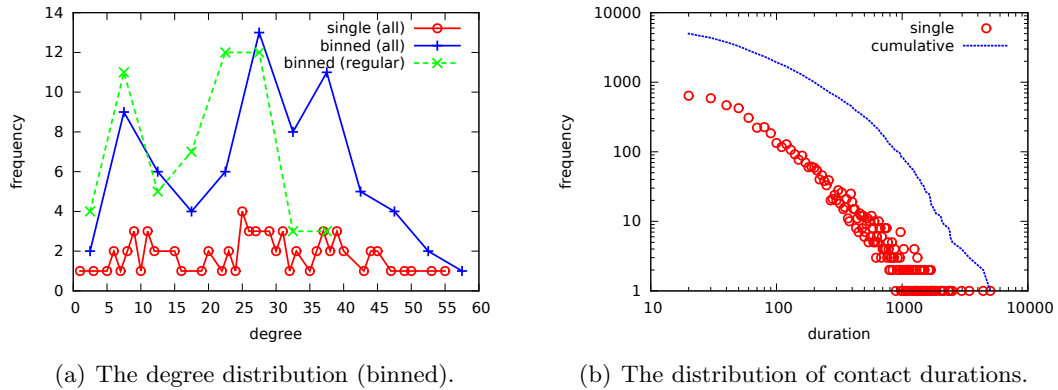


Figure 4.2: The degree distribution (left) in three versions: The actual distribution (single) and distributions over degree bins of breadth five, including either contacts between all participants or only those between regular participants. The distribution of the contact durations (in seconds) is shown on the right, together with its cumulative version. The distributions are observed in  $LWA[\geq 0]$ .

#### 4.5.2 Workshops

In this section, we approach Research Question RQ2 by discussing community structure that can be detected in the interaction network, and by comparing it to the partition of the participants into the four special interest groups ABIS, IR, KDML, and WM. LWA 2010 was a workshop event of these four special interest groups. Table 4.2 shows the number of participants in each workshop and the number of organizers and regular participants. Setting  $i$ , the minimum duration for contacts in the networks  $LWA[\geq i]$ , to  $i = 5$  or  $i = 10$ , and removing isolated nodes reduces the number of participants with at least one contact in each workshop slightly. Setting  $i$  to  $i = 30$  reduces that number drastically. The same effect is visible for the organizers and regular participants. Here, we see a pronounced difference between these two groups: All organizers are still connected in  $LWA[\geq 10]$ , and in  $LWA[\geq 30]$  half of the organizers are still connected to some other participant. Among the regular participants, however, about 20% are isolated in  $LWA[\geq 10]$ , and in  $LWA[\geq 30]$ , only 20% are still connected to at least one other participant. This is again evidence for the special role organizers played at the event.

Since the special interest groups capture common research interests and along with them also personal acquaintance, it is reasonable to expect that the set of participants is naturally clustered accordingly, at least to a certain extent. To get an impression on that clustering, we compute the community alignment metrics of Scripps et al. [2007b], recalled here in Definition 2.5. For four thresholds  $i$ , the results are shown in Table 4.3. Overall, both statistics are relatively low. For  $LWA[\geq 0]$ , we see that less than half of all edges run between participants of the same workshop. The assumption

Table 4.2: Distribution of the non-isolated nodes over the four workshops and over regular participants and organizers in different contact networks of LWA 2010.

network	ABIS	IR	KDML	WM	regular	organizer
LWA[ $\geq 0$ ]	7	11	37	16	59	12
LWA[ $\geq 5$ ]	7	11	34	15	55	12
LWA[ $\geq 10$ ]	6	9	32	13	48	12
LWA[ $\geq 30$ ]	0	3	10	5	12	6

Table 4.3: Community Alignment based on the four tracks in the contact networks for all participants and for regular (non-organizer) participants only. Isolated nodes are ignored.

network	all		regular	
	p	q	p	q
LWA[ $\geq 0$ ]	0.45	0.72	0.39	0.76
LWA[ $\geq 5$ ]	0.58	0.70	0.58	0.76
LWA[ $\geq 10$ ]	0.68	0.68	0.67	0.78
LWA[ $\geq 30$ ]	0.64	0.65	0.80	0.84

that this would be due to the influence of the organizers, can be rejected by observing the alignment metrics for the subgraph containing only regular participants, where the share is even lower. Increasing the threshold to  $i = 5$  or  $i = 10$  (and ignoring the resulting isolated nodes), raises the share of edges within the workshop communities ( $p$ ).<sup>12</sup> Thus, more of the longer talks actually happen between participants of the same workshop. Still, even for  $i = 10$  one third of the edges connects participants from different workshops. We can interpret these values as evidence that the interdisciplinary nature of the event (including shared sessions and a social gathering at the poster session) or the relatedness of the topics in the workshops influence the choice of conversational partners.

To investigate the situation more in-depth, next, we count the number of contacts from participants of one workshop to participants in any workshop, and we compare that number to the expected value for that quantity if conversational partners were chosen unbiased from the workshop they registered for. Table 4.4 shows these fractions. We can see clearly that conversations are biased towards members of one's own workshop. Restricting the set of contacts to those between regular participants changes only little, mainly the ratios between KDML and some other workshop change. This is due to the fact that eleven of the twelve organizers registered for KDML. The

<sup>12</sup>The value of  $q$  decreases slightly for all participants and increases slightly for regular participants, but overall does not change much except for the highest considered threshold  $i = 30$ .

Table 4.4: Actual versus expected number of contacts between workshop members: For each workshop  $A$  (row), shown are for each workshop  $B$  (column) the ratio of the average number of contacts from a member of  $A$  to members of  $B$  and the expected value for that quantity (assuming the same number of contacts from members of  $A$  and an unbiased selection of dialog partners).

workshop	all				regular			
	ABIS	IR	KDML	WM	ABIS	IR	KDML	WM
ABIS	<b>4.36</b>	0.52	0.63	0.93	<b>4.12</b>	0.49	0.52	0.88
IR	0.46	<b>2.89</b>	0.80	0.51	0.47	<b>3.56</b>	0.52	0.51
KDML	0.28	0.41	<b>1.49</b>	0.62	0.38	0.40	<b>1.57</b>	0.82
WM	0.46	0.28	0.68	<b>2.58</b>	0.44	0.27	0.56	<b>2.46</b>

remaining organizer registered for IR, and we see, as expected, that the ratio for contacts between members of IR rises when the organizers are removed.

Finally, we compare the community allocation induced by the workshops to communities detected by graph mining algorithms. We compare the respective allocations using modularity (see Definition 2.6). We apply two heuristics that optimize community allocations towards modularity: *Infomap* [Rosvall and Bergstrom, 2007] and *Louvain Clustering* [Blondel et al., 2008], which were shown to perform well [Lancichinetti and Fortunato, 2009]. In the previous section, we saw that setting the threshold  $i$  on the minimum duration of contacts to  $i = 5$  is a good compromise between removing short contacts and keeping most of the participants in one connected component. Therefore, we use  $\text{LWA}[\geq 5]$  to detect communities. Furthermore, since modularity can be computed on weighted graphs, we can use a weighted version of  $\text{LWA}[\geq 5]$ , where each edge is weighted with the the sum of the durations of all contacts between the respective participants.

The modularity computed on the community allocation induced by the workshops is  $\text{MOD} = 0.20$ . The two heuristics *Louvain Clustering* and *Infomap* find different, yet similar clusterings with higher modularity,  $\text{MOD} = 0.27$  in both cases. This indicates that the workshop partitioning indeed covers some aspects of the community formation (otherwise modularity would be zero), yet it does not explain the structure fully. As before, we check whether the organizers in particular hinder a workshop-induced community structure, and we therefore compute community allocations and modularity using only the regular participants. We yield  $\text{MOD} = 0.26$  for the workshops and  $\text{MOD} = 0.33$  and  $\text{MOD} = 0.34$  for *Louvain Clustering* and *Infomap*, respectively. As for the full set of participants, the allocations computed by the two heuristics yield a higher modularity than the one induced by the workshops. Thus, we can conclude, that not only the organizers but many other participants as well connect to their peers in other special interest groups.



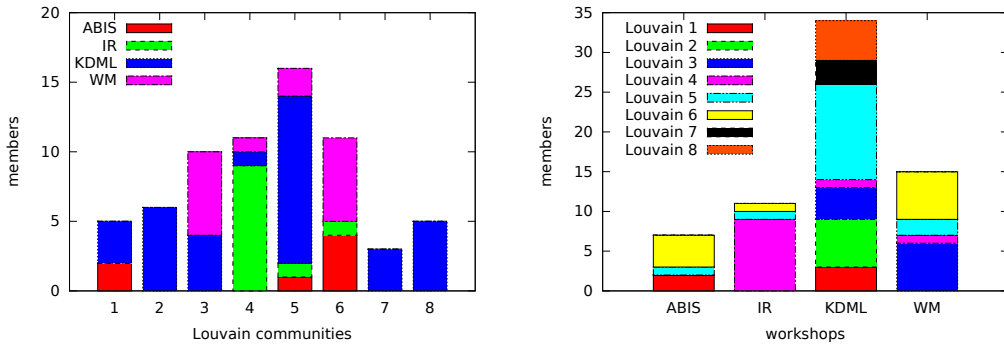
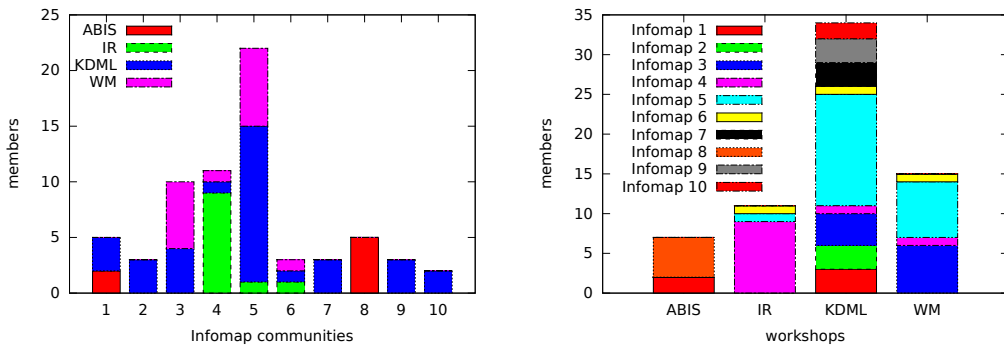
(a) Workshops in *Louvain Clustering* communities. (b) *Louvain Clustering* communities in workshops.(c) Workshops in *Infomap* communities.(d) *Infomap* communities in workshops.

Figure 4.3: Distribution of the members of the four special interest groups (workshops) across the communities found by the community detection heuristics *Louvain Clustering* and *Infomap*; and vice versa the distribution of the mined communities across the four workshops. Communities are computed on  $LWA[\geq 5]$  where the edges are weighted by the total contact duration. Each color corresponds to one single community.

Figure 4.3 shows how the mined communities and the workshop based allocation overlap (for the graph including all participants – the situation on the graph between only regular participants is similar). Both heuristics detect more than just four communities (like the workshop based allocation), and many communities contain members of more than one special interest group (Figures 4.3(a) and 4.3(c)). Thus, they are not simply sub-communities of the workshops. Yet, there are several communities that contain only members of a single workshop, and almost all communities have one workshop that dominates among its members. Each special interest group is distributed over several communities (Figures 4.3(b) and 4.3(d)). The most extreme example is the largest workshop, KDML, which contains members from seven communities according to the result of *Louvain Clustering* and from nine communities according to *Infomap*.

Overall, we can conclude, that community structure can be found in the contact graphs. Part of that structure can be explained by the clustering into the four workshops, yet communities transcend that clustering. Communities have been formed within the workshops as well as among participants of different special interest groups. While on average, more conversations are directed towards members of the own group, there is also a considerable amount of interaction with members of other groups.

### 4.5.3 Peer Groups

In the last part of our analysis, we investigate participants and their role in the contact network, and we compare these positions on the level of peer groups, thus attending to our third research question (RQ3). For that purpose, we group all participants in peer groups, either by their academic status or by their function at the conference (organizer or regular participant).

#### Centrality

Table 4.5 displays the average and median values of several graph node centralities (see Section 2.1.2) of  $LWA_{\geq 5}$ , aggregated (average  $\mu$  and median  $\tilde{x}$ ) by academic position, as well as separately by conference organizers and regular conference participants. We set the threshold to  $i = 5$ , since we saw in the previous section that a lot of brief contacts are indeed filtered, yet still almost all of the participants are still connected to one large component. Note, that while the categories referring to academic status are disjoint (the category *other* includes all participants that do not fit one of the other four) organizers and regular participants both include participants from all the academic status categories.

A first observation is that the organizers have significantly higher scores in all measures under observation. In the considered scenario this is highly plausible due to the nature of an organizer's job during a conference – which in the case of LWA 2010 also included the supervision and maintenance of the RFID-experiment and of Conferator. It is, however, interesting to note that this effect is visible even though we used a minimum threshold of  $i = 5$  minutes for contact durations and, thus, filtered short conversations that one might assume to be typical for organizational purposes.

A second observation is that the distribution of the centralities within one group is often skewed, which can be seen by comparing the mean and the median of each measure. The difference is particularly strong for the betweenness of students, where the average of  $\mu = 70.62$  is far from the median value  $\tilde{x} = 15.19$ . In fact, the high average is due to one very active student who had a betweenness score of  $\text{bet} = 219.98$  alone. Because of the event's small size and the even smaller, status-based subsets, single individuals can strongly influence the scores for their groups. Comparing median and mean for the other groups and measures, we see that the median is almost always below the average; sometimes far below (e.g.,  $\text{str}_{\#}$  and  $\text{str}_{\Sigma}$  for professors and postdoctoral researchers). The exception to that rule is betweenness for postdocs. This

Table 4.5: Group size and average ( $\mu$ ) and median ( $\tilde{x}$ ) graph centralities per academic position, as well as for organizers and regular participants in LWA[ $\geq 5$ ]: degree deg, strengths  $\text{str}_\#$  (where edges are weighted by the number of contacts they represent) and  $\text{str}_\Sigma$  (where edges are weighted by the sum of the durations (in seconds) of all contacts between the respective nodes), and betweenness bet.

status	size	deg		str $_\#$		str $_\Sigma$		bet	
		$\mu$	$\tilde{x}$	$\mu$	$\tilde{x}$	$\mu$	$\tilde{x}$	$\mu$	$\tilde{x}$
Professor	12	<b>9.3</b>	<b>7</b>	<b>20.8</b>	11	<b>14,966</b>	8,801	58.23	45.57
Postdoc	9	8.2	6	15.6	9	10,140	5,560	<b>74.87</b>	<b>83.98</b>
PhD st.	35	8.1	<b>7</b>	18.5	<b>15</b>	11,772	<b>9,212</b>	49.34	26.53
Student	7	7.4	6	16.9	<b>15</b>	8,913	8,227	70.62	15.19
Other	8	5.9	3	12.6	7	8,487	5,152	37.29	15.61
Organizer	12	<b>12.7</b>	<b>12</b>	<b>31.3</b>	<b>29</b>	<b>20,325</b>	<b>19,166</b>	<b>90.34</b>	<b>77.43</b>
Participant	59	7.1	6	14.9	12	9,648	6,666	47.59	26.53

means that in almost all cases, there are a few individuals with scores far above the respective group’s average. This observation motivates the use of social conference tools, like Conferator, that could assist the majority of participants in initiating contacts to their community and to persons of interest. The observation also suggests that personal traits, beyond the academic status are influential factors for a participant’s position within the contact network.

With the limitation of small group sizes in mind, we still notice striking differences between the academic status groups. First of all, on average, the group of professors has the highest degrees and strengths. They are followed by the PhD students (who have the highest median values for these measures) and the post doctoral researchers. The students have lower average degree and strength when edges are weighted with their sum of all contact durations ( $\text{str}_\Sigma$ ), while the strength in the graph counting the contact frequencies ( $\text{str}_\#$ ) is between the values for postdocs and PhD students. This indicates that on average, students had as many contacts as other participants, however, with fewer conversational partners and for shorter talks. We attribute this phenomenon to the fact, that students are less established in their scientific communities than scientists in higher academic positions and usually have little conference experience. On the other hand, the observed differences between students and postdocs or PhD students are not particularly strong. The average values are mostly comparable to those of other groups and the median values are comparable and sometimes even higher than those of others.

Another aspect is illustrated by the betweenness (bet) scores: Relatively to the other groups, a lot of shortest paths of LWA[ $\geq 5$ ] run through nodes of postdocs.

Considering the typical structure of research institutes, where usually one professor supervises several postdocs, who again each supervise several PhD students, postdocs are the connection between professors and postgraduates. They thus assume the role of gatekeepers in their institutions, and we hypothesize that this effect is reflected in the betweenness scores in the LWA contact network.

### Community Roles

The assignment of roles to nodes in a network is a classification process that categorizes the players based on their position in the graph. In the following, we investigate how participants in their explicitly given roles, like the academic position or the job (organizer), fill the implicitly determined, graph structure-based roles. Therefore, we apply the community role classifier of Scripps et al. [2007a], recalled here in Section 2.1.2, to the graphs  $LWA[\geq 0]$  through  $LWA[\geq 15]$ , and we determine – under the assumption, that longer contacts indicate more serious, research related discussions – how the filtering of short contacts changes community roles.

The community role assignment for participants depends on their node degree and the community metric `rawComm`, which measures how many communities a node connects. The `rawComm` can be computed either based on a given community assignment or using a statistical approach, relying on estimates for the community alignment metrics  $p$  and  $q$  from Definition 2.5. Since we have observed above that the community assignment induced through the special interest groups, does not fully explain the community structure in the contact network, we choose the probabilistic approach, and we use the special interest groups only to estimate the parameters  $p$  and  $q$  in the respective graphs. For the role assignment according to the rule in Formula 2.1, two thresholds need to be selected:  $s$  for the degree and  $t$  for the `rawComm` measure of a participant. We normalize both metrics to the interval  $[0, 1]$  and set the thresholds to  $s = t = 0.5$ , the same setting as in [Scripps et al., 2007a].

The first immediate finding is that in none of the graphs  $LWA[\geq 0]$  through  $LWA[\geq 15]$ , any participant is ever classified as Big Fish, that is, whenever a node has a high degree, it also has a high `rawComm` score. We attribute this peculiarity to the fact that the purpose of social interaction at conferences often is the exchange of ideas with colleagues outside the own work group. Especially during LWA 2010, participants were encouraged to engage in interdisciplinary dialogue, for example, by including several joint sessions in the schedule and a combined event of social dinner and poster session.

The distribution of participants among the three remaining nodes is depicted in Figure 4.4. For instance, Figure 4.4(a) displays for each academic position, as well as for organizers and for all participants together, the percentage of participants in the respective groups that have been classified as Ambassador. Figures 4.4(b) and 4.4(c) show these shares for the roles Loner and Bridge. For example, we can see in Figure 4.4(a) that in  $LWA[\geq 15]$ , about 42% of the professors (5 out of 12) have been classified as Ambassador. Half of the professors have been classified as

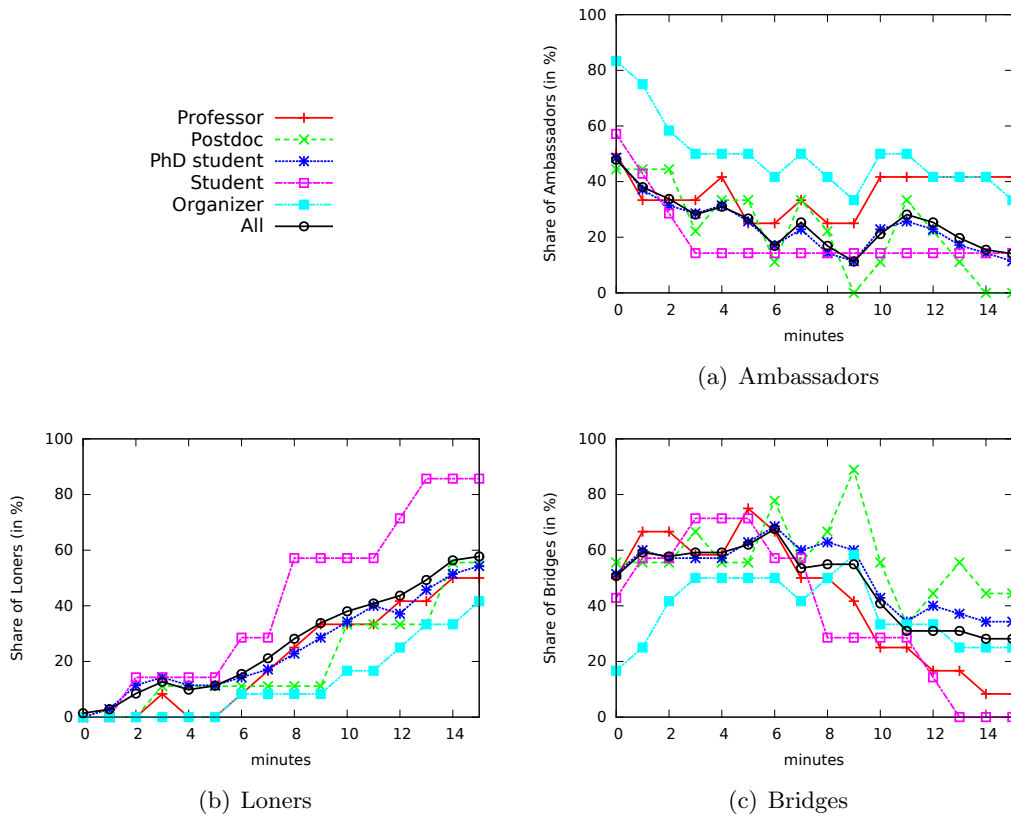


Figure 4.4: Fraction of participants that assume the community roles Ambassador, Bridge, or Loner, for different thresholds  $i$  in  $LWA[\geq i]$ . In each plot, shown is the fraction of participants with the same academic status that assume the particular role. Additionally shown is that fraction among all participants and among the organizers.

Loner (Figure 4.4(b)) and only one professor has been assigned the role Bridge (8%, Figure 4.4(c)). The sizes of each group can be found in Table 4.5.

All curves in Figure 4.4 fluctuate: Since with rising the threshold  $i$  on the contact duration, the graphs get more sparse, the community alignment metrics change and with them, the rawComm metric. Still, there are several clearly visible tendencies. In all three diagrams, the fractions of PhD students is very close to the fraction of all participants (All). The simple reason for that is, that PhD students are the majority within the conference population and therefore dominate the general behavior. For all groups, the share of Loners rises with higher threshold  $i$ , which is also plausible as all participants lose edges and thus degree and rawComm.

Many of the organizers start out as Ambassador. This is again consistent with their job description. However, filtering out short contacts and thus the typical quick organizational conversations, the relevance of the organizers decreases. More and

more organizers become Bridges and even Loners with higher  $i$ ; in the last graph,  $LWA[\geq 15]$ , they are almost equally distributed among the three roles. One should keep in mind, that the group of organizers contains persons in all academic positions. Therefore, after filtering out most of the contacts that, presumably, are due to their organizational work, the organizers act mainly in their respective role as conference participants, which might explain the stronger fluctuations in the right part of the curve.

Very consistent with the findings on graph centralities, described above, is the role distribution among the students. While in the first graphs, where short contacts dominate the longer ones, some of them are classified as Bridge or Ambassador, they quickly disappear from those roles and are classified as Loner, with the exception of one particularly active student who is classified as Ambassador in every graph..

Compared to the PhD students, the fractions of postdocs are, with few exceptions, higher for the role Bridge and lower for Loner. This is again consistent with the previous observations concerning the graph structure measures. Due to their greater experience postdocs seem to have more access to colleagues in other communities. Finally, the curve of the professors in the role Ambassador shows the largest deviation from the mainstream. While in that role all other groups' fractions decrease, that of the professors remains stable at about 42%.

In summary, we observe that the chosen method of role assignment is, although dependent on few individuals, still somewhat correlated with the participants' academic positions. Several professors tend to assume the most active role (Ambassador), students are more often classified as Loner and PhD students and postdocs are somewhere in between.

## 4.6 Conclusion

In this chapter, we have presented an in-depth analysis based on data from the dissemination phase of the publication life cycle, collected during the LWA 2010 in Kassel, in October 2010, using the social conference guiding system Conferator. While these results are specific to LWA 2010, the techniques we applied are universal and they can similarly be used on other research conferences (workshop events like LWA or multi-track conferences) as well.

We investigated graph properties of the observed contact networks and found both similarities and differences to a previous experiment at the Hypertext conference of 2009, answering Research Question RQ1.

Regarding our second research question (RQ2), we could show that there is a detectable community structure in the investigated face-to-face networks. Comparing community allocations that were mined from the contact graph, to a clustering into the four special interest groups that attended the conference, we saw that they are not consistent. The special interest groups do induce a certain community clustering, and on average, members from the own community are more often chosen as conversational

partner than others. However, conversations transcend the special interest groups, and most of the mined communities (which reflect the clustering on the observed contact network) are actually comprised of members from more than one special interest group. A possible interpretation of these results is that the LWA as a joint workshop event for four groups (rather than four individual events) is a successful concept, as interdisciplinary exchange indeed takes place.

Finally, addressing Research Question RQ3, we could observe that typical graph centrality measures are not equally distributed over the participants, grouped by their academic status. A similar behavior can be seen from roles that can be assigned based on a participant's position in the face-to-face contact networks. In both cases, professors are the most active group, while many students are not as well connected as other attendees. This fits well to the intuition that participants with longer academic experience are the ones with the larger network. A possible interpretation for conference organizers is that they should consider steps to encourage interaction between those who are long-time members already and those who only recently joined the community.

Our findings motivate the use of tools like Conferator, that support networking processes particular for those participants who are relatively new to the community. LWA is a rather small conference, and many participants have known each other already before the event. We can assume that at larger events, where participants come from more different disciplines, participants would profit even more from Conferator's features.

#### 4.6.1 Future Research

After the original publication of the analyses presented here [Atzmueller et al., 2012b], both Conferator, as well as this particular line of research – investigating the dissemination phase of the publication life cycle – have been continued. Conferator has been further developed by others and became part of the open source framework Ubicon.<sup>13</sup> A new feature, providing utility to participants is the *acquaintomatic* [Atzmueller et al., 2012a] – a user recommender system.<sup>14</sup>

Through deployments at further conferences, others have acquired data from more diverse conferences and conducted further analyses. For instance, Macek et al. [2012] continued and extended the analyses presented here, using data from Hypertext 2011. Kibanov et al. [2013] investigated how communities change over time during a conference, as well as their stability and predictability. Scholz et al. [2014] observed that face-to-face contacts can help improve a topic-model-based predictor for talk attendance, while in [Scholz et al., 2013], it was demonstrated that data from online social networks can help predict face-to-face contacts during a conference. Kibanov et al. [2015] used thresholds for the duration of face-to-face contacts to generate contact networks of different sparsity levels. They found structural similarities to

<sup>13</sup><https://bitbucket.org/ubicon/ubicon>

<sup>14</sup>The version of Conferator used to gather the data for this chapter, did not contain the *acquaintomatic*, thus, our analysis is unbiased from user recommendations.

online social networks to be stronger for thresholds of three or five minutes rather than for unrestricted contact networks. This observation is evidence for the importance of longer contacts that we have presumed in our analysis here. Finally, Atzmueller et al. [2016b] investigated how groups form and dissolve during a conference and particularly during coffee breaks. Like in this chapter, they used data that was gathered during LWA 2010.

By and large, the analysis of a conference's participants' interactions has proven to be a fruitful area, and the results can help participants in their networking efforts, as well as organizers plan the event. Further directions for future research include the evaluation of recommender systems for face-to-face contacts and the analysis of other research communities (the conference in this chapter and those in the work mentioned above all belong to the field of computer science).



## **Part II**

# **Analyzing the Usage of Scholarly Social Bookmarking**



## Chapter 5

### Analyzing Scholarly Publication Management



Social tagging systems have established themselves as an important part in today's web and dedicated scholarly tagging systems support researchers in their daily work with publications. Hence, they have attracted the interest of our research community in a variety of investigations. Several aspects of social tagging systems have been discussed and assumptions have emerged on which our community also builds their work. Yet, testing such assumptions has been difficult due to the absence of suitable usage data in the past. In this work, we thoroughly investigate and evaluate four aspects about tagging systems, covering social interaction, retrieval of posted resources, the importance of the three different types of entities, users, resources, and tags, as well as connections between these entities' popularity in posted and in requested content. For that purpose, we examine live server log data gathered from the real-world, public social tagging system BibSonomy. Our empirical results paint a mixed picture about the four aspects. While for some, typical assumptions hold to a certain extent, other aspects need to be reflected in a very critical light. Our observations have implications for the understanding of social tagging systems, and the way they are used on the web. The results presented in this chapter have also been published in [Doerfel et al., 2014c] and [Doerfel et al., 2016b].

### 5.1 Introduction

With this chapter, we continue our journey through the publication life cycle to the usage phase. At the same time, this is the first chapter that belongs to the social bookmarking theme of this thesis. In the usage phase of the publication life cycle, a publication is (hopefully) read by others. Due to the large amount of available literature, readers have to find ways to organize the literature they plan to read, have read, or consider citing in their own work. With BibSonomy (Section 2.3.2), CiteULike,<sup>1</sup> Connotea,<sup>2</sup> or Mendeley,<sup>3</sup> social bookmarking systems have been created

---

<sup>1</sup><http://www.citeulike.org/>

<sup>2</sup><http://www.connotea.org/>

<sup>3</sup><https://www.mendeley.com/>

that are dedicated particularly to the management of publications. They offer a means of organizing information, and they have established themselves as an alternative to more traditional resource directories. This chapter and the next focus on the core features of tagging systems: tagging, navigation on the folksonomy structure, and retrieval. We investigate five aspects of such systems – four that apply to any bookmarking systems in this chapter; a fifth is specific to scholarly bookmarking and will be the subject of Chapter 6.

Social tagging systems in general have attracted the interest of our research community for over a decade [Mathes, 2004, Golder and Huberman, 2006]. Significant advances have been made with regard to our understanding about the emergent, individual and collective processes that can be observed in such systems; useful algorithms for retrieval have been developed that exploit the rich fabric of links between users, resources, and tags in social tagging systems for facilitating information organization, search and navigation; and further work has focused on the extraction or stabilization of emergent semantics (see Section 2.3.3). While this line of research has significantly increased our ability to describe, model, and utilize social tagging systems, our community has also built their work on certain assumptions about how these systems are used. However, whether – and to what degree – these assumptions hold, is still an open research question. In the literature, arguments and evidence regarding the usage of tagging systems have been discussed controversially and researchers have argued for and against them, providing thus all the more reason to evaluate them on real-world usage data. Only a few studies have actually analyzed user behavior in social tagging systems to better investigate these research questions, either by conducting user surveys (e.g., Heckner et al. [2009]) or by tapping into the rich corpus of tagging data (i.e., the posts) that is available on the web (e.g., Cattuto et al. [2007]). However, such studies come with certain limitations such as self-reporting biases or the lack of detailed usage data revealing how users actually request information. In this chapter, we overcome these drawbacks by presenting and thoroughly investigating a detailed usage log of the popular real-world, open social tagging system BibSonomy. We thus provide evidence from actual user behavior to shed light on a series of questions from related work regarding the usage of a tagging system.

**Research Questions.** The research questions in this chapter are aligned along the following four controversial aspects about the usage of social tagging systems:

- (RQ1) *The social aspect:* Tagging systems are supposed to be used collaboratively to tag and share resources. *We investigate to which degree such sharing actually happens and discuss evidence for the interest of users in the content of others.*
- (RQ2) *The retrieval aspect:* The main activities in a tagging system are storing resources and retrieving them later (using the assigned tags). *We investigate whether and when users retrieve their resources.*

- (RQ3) *The equality aspect:* In the folksonomy model (Section 2.3.1), users, resources, and tags are modeled as equally important sets of entities. *We investigate whether they are indeed equally important in navigation or whether one of the three types of entities is preferred for retrieving and browsing content.*
- (RQ4) *The popularity aspect:* The popularity of users, tags, and resources in posts is often seen as an indicator of importance – for example in tag clouds where frequent tags have large font sizes to gain the users’ attention and to be easily accessible by a mouse click. *We investigate whether popularity in posts is matched by popularity in retrieval.*

For each of these four aspects one could formulate a (naive) assumption about tagging systems:

- social tagging systems are (as their name suggests) social;
- users do retrieve their resources after they have stored them;
- users, resources, and tags are equally important for navigation; and
- popularity in posts implies popularity in requests.

These assumptions are very plain statements and we do not expect to find them confirmed just like that. However, as they reflect beliefs about tagging systems (evidence from the literature for each aspect is presented in the according subsections of Section 5.5), it is worth investigating to what degree they actually do hold. In this chapter, we investigate them in the use case of publication management.

**Contributions.** This chapter makes contributions on two levels:

1. *Methodical:* We identify a number of aspects and illuminate a way towards investigating them with log data.
2. *Empirical:* We investigate a number of research questions regarding user behavior in a social tagging system by testing them with actual log data and we report the results exemplarily for the scholarly social tagging system BibSonomy. The study is enabled by data generated in the usage phase of the publication life cycle.

Overall, our findings are relevant for researchers interested in user behavior and modeling in the context of social tagging systems and their adoption, as well as to system engineers interested in improving the utility and usefulness of social tagging systems, particularly for publications, on the web.

**Limitations.** While our findings are limited to the scholarly bookmarking system BibSonomy, our method of examining social bookmarking systems is general. The approach itself is independent of the resources that are bookmarked (in this case publications); and it can well be applied to other social tagging systems to investigate these aspects in different contexts and to test assumptions like the ones above. We discuss and speculate about influences on the behavior in tagging systems of other (non-scholarly) contexts in Section 5.2.

**Structure.** In the next Section, we introduce the use case BibSonomy and we discuss issues of generalizability of our results. After the discussion of related work in Section 5.3, we describe the BibSonomy datasets in Section 5.4. We then turn our attention to studying and evaluating the aforementioned four aspects on social tagging in Section 5.5, approaching each of the above research questions individually. Finally, Section 5.6 concludes the chapter.

This chapter is based on the publications [Doerfel et al., 2014c], [Doerfel et al., 2014b], and [Doerfel et al., 2016b]. Particularly the latter contains all results published here. For this thesis they have been slightly rearranged.

## 5.2 The Use Case BibSonomy

We have introduced the publication management system BibSonomy in Section 2.3.2. In this (and the next) chapter, we will use BibSonomy as our use case to study the behavior of users in a scholarly social bookmarking system. In this section, we describe some aspects of the navigation in BibSonomy, thus, the framework within which users can interact with the system. Then, we discuss the generalizability of our approach.

### 5.2.1 BibSonomy

As a social bookmarking system the two central activities for users are *storing* (posting) and *retrieving* resources. Publications and web links can be stored in BibSonomy by entering the respective data manually into input masks. To simplify this process, BibSonomy offers bookmarklets and browser add-ons.<sup>4</sup> When users visits a web page they want to bookmark directly, or which contains a publication reference they want to store, they can click a button and are forwarded to the input mask that is already filled. Users only have to add tags and are referred back to the original page after they have created the post. BibSonomy also supports mass imports of both publications and web links (e.g., BIBTEX files or browser bookmark collections).

#### Retrieving Content

In our context, we consider any request to a page with user-generated content. This includes post lists or resource pages, and excludes, for example, settings or help pages.

---

<sup>4</sup><http://www.bibsonomy.org/buttons>

Every page in BibSonomy is identified by a unique URL, and thus we can tell from the system's request logs which retrieval activity a user has chosen. The URL schema of BibSonomy is described in its help pages.<sup>5</sup> BibSonomy offers for example the following options to query for posts: A user can request to see all posts with one or several tags,<sup>6</sup> all posts of a specific user<sup>7</sup> or group,<sup>8</sup> and also use a combination of user and tag restrictions.<sup>9</sup> Figure 5.1(a) shows the bookmarks and publications of the user "hotho" with the tag restrictions "web" and "mining". Users can also use a full text search for retrieval.<sup>10</sup> In BibSonomy, the full text of a post includes its metadata, owner, and tags.

For each resource, BibSonomy has a page that lists its tags and users from all posts.<sup>11</sup> Publication posts have a *details* page<sup>12</sup> (for an example see Figure 5.1(b)), that shows the metadata of the publication (as entered by the user who created the post) and offers export options. While user pages that are restricted by a tag (mentioned above) and details pages of posts each combine two folksonomic entities – a user and a tag in the first case and a resource and a user (the post's owner) in the second –, a combination of tag and resource does not exist.

Posts of bookmarked websites can also contain metadata (like a description of the website), but requests to a bookmark post are usually directly to the bookmarked website and thus external requests. For example in a post list, the title of each publication post links to the post's details page, while the title of each web page post links directly to that page. Such requests are not recorded in BibSonomy's server logs, and therefore, we must restrict some experiments exclusively to publication requests.

### Further Features

In BibSonomy, users can form groups or declare friendships to other users. Both friendships and groups are used in the visibility concept of posts. We will take a look at both in our discussion of the social aspect in Section 5.5.1. BibSonomy offers many further features like discussion forums that exceed the usual tagging system functionality. Therefore, such features have been excluded from our experiments.

Due to its high rank in search engines, BibSonomy is a popular target for spammers. Spammers are users who store links to advertisement sites to increase their visibility on the web. For regular users (non-spammers) it becomes harder to retrieve the relevant content. BibSonomy uses a learning classifier [Krause et al., 2008] as well as manual classification by the system's administrators to detect spam. Users can be classified as spammers or non-spammers based on their profile and their activities in the system. In all experiments, we only used data of users that have been classified as non-spammers.

---

<sup>5</sup>[http://www.bibsonomy.org/help\\_en/URL-Syntax](http://www.bibsonomy.org/help_en/URL-Syntax)

<sup>6</sup>e.g., <http://www.bibsonomy.org/tag/web+mining>

<sup>7</sup>e.g., <http://www.bibsonomy.org/user/hotho>

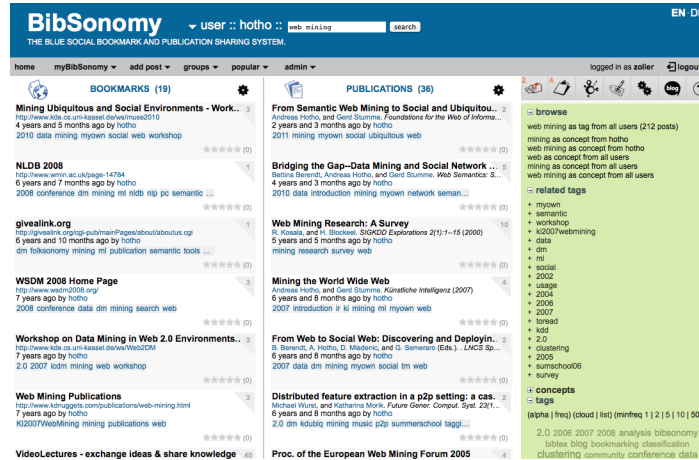
<sup>8</sup>e.g., <http://www.bibsonomy.org/group/kde>

<sup>9</sup>e.g., <http://www.bibsonomy.org/user/hotho/web+mining>

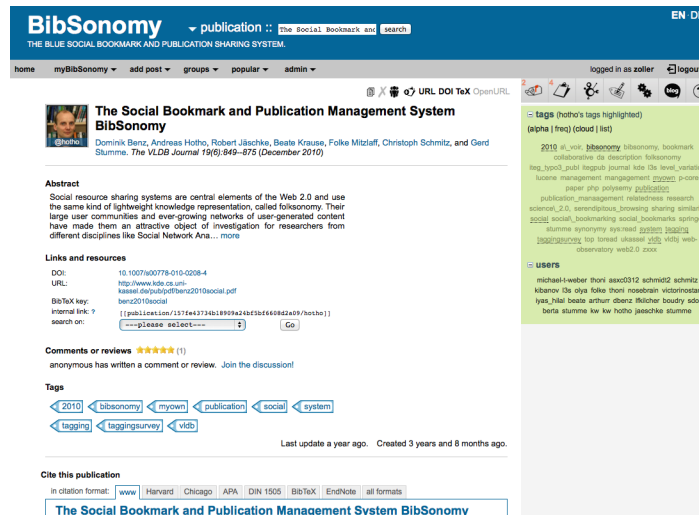
<sup>10</sup>e.g., <http://www.bibsonomy.org/search/Semantic+Web>

<sup>11</sup>e.g., <http://www.bibsonomy.org/bibtex/157fe43734b18909a24bf5bf6608d2a09>

<sup>12</sup>e.g., <http://www.bibsonomy.org/bibtex/157fe43734b18909a24bf5bf6608d2a09/hotho>



(a) user page with tag restrictions (tags: “web” and “mining”)



(b) publication details page

Figure 5.1: Screenshots of BibSonomy’s web interface: (a) illustrates the typical resource list layout, showing both bookmarks and publications side by side, and (b) displays a publication details page, which lists all information entered by a user for a specific publication. The screenshots were taken in September 2014. In late November 2014, a new layout has been introduced.

### 5.2.2 Generalizability

In the following sections, we will investigate four aspects of social tagging, using the system BibSonomy as our showcase. While our findings in this chapter are limited to BibSonomy, our approach is directly applicable to other tagging systems, and we briefly discuss some aspects of such a transfer here. Where possible, we compare our



results on BibSonomy to results from previous studies. However, to the best of our knowledge, for no other social bookmarking system, such web log analyses have been published yet. Like shown in the user study by Heckner et al. [2009], different tagging systems exhibit different characteristics (in their case regarding the users' tagging motivation). In the following, we speculate about possible influences and results in other tagging systems.

One influence for the behavior of users is surely the system's *degree of openness*. In contrast to open, publicly available systems, company-internal systems, like Dogear [Millen and Feinberg, 2006], can impose certain requirements on their users, like the use of real names instead of pseudonyms or boundaries for the tags and resources in the system. For example, the knowledge whose resources one browses could be a strong influence for the social behavior of sharing and visiting. Indeed, we will see similarities but also pronounced differences in the usage behavior in BibSonomy compared to that in Dogear in our investigation of the social aspect in Section 5.5.1 and also in the popularity aspect in Section 5.5.4.

Another factor is surely the *type of resources* that are bookmarked. Heckner et al. [2009] have shown that motivations for tagging (sharing or personal information management) were different in the systems YouTube (resources are videos) and Flickr (images) compared to Delicious (web links) and Connotea (publication references). A major difference between those two pairs of systems is that links and publication references are taken from other, already available sources, whereas images and videos are often published in the respective system for the first time. With regard to the social aspect, we therefore would expect to find that similar studies on Delicious and Connotea, compared to BibSonomy, would yield similar results, because BibSonomy allows users to tag web links (like Delicious) and references to publications (like Connotea). On the other hand, we can speculate that systems like YouTube and Flickr would show different results, for instance, a much stronger interest in the content of other users.

The *age* of the system is another influence. Previous log file analyses on other systems [Millen and Feinberg, 2006, Damianos et al., 2007, Millen et al., 2007] report results from periods of eight, ten, and twelve months respectively, shortly after the systems' creation in 2005. In contrast to that, our log dataset covers a period of six years (cf. Section 5.4).

Finally, the *navigation concept* and the *graphical user interface* can play a role. BibSonomy offers the typical folksonomy navigation by always presenting users, resources, and tags as linked entities. However, different tagging systems may make different design choices, for instance, regarding the visibility and accessibility of individual entities.

To investigate these questions further, one would have to conduct experiments on logs of other systems as well. However, the bottleneck, here, is the availability of such datasets. Therefore, our study is a first step towards analyzing user behavior using log files. We encourage other researchers and webmasters of tagging systems to conduct

similar studies, using the here presented methods, on their tagging systems and to compare their results to ours.

## 5.3 Related Work

In this section, we discuss related literature on the investigation of tagging systems and log file analysis in general. Further related work, that is specifically relevant to individual aspects, will be discussed in greater detail later in the corresponding context in Section 5.5. Since we have already mentioned some research on various aspects of social bookmarking systems in Section 2.3.3, here, we focus on work directly relevant to this study. For details on BibSonomy, the subject of the investigations in this chapter, we refer to Section 2.3.2 and to the previous section.

### 5.3.1 User Surveys and Post Analysis

Abrams et al. [1998] already discussed the management of website bookmarking long before the rise of social bookmarking on the web using a user survey and bookmark files from participants. Their results showed that users are motivated to share bookmarks (still via email back then), as well as to retrieve them later. Heckner et al. [2009] conducted a survey of tagging systems (namely Flickr, YouTube, Delicious and Connotea) with 142 users regarding their motivations. It showed that there are mainly two motivations for users to post content: sharing resources with others and storing them primarily for personal retrieval later on. The strength of these motivations varies from system to system.

Using the post data of tagging systems, several studies analyzed aspects of posting behavior, for instance, the distributions of users, resources, and tags in posts [Cattuto et al., 2007], or the identification of different types of users – categorizers and describers – regarding their choice of tags [Strohmaier et al., 2010]. However, these studies did not use log data for their analysis to explore the actual retrieval behavior. A review of social tagging, regarding a variety of diverse aspects of such systems – including vocabulary, structure, visualization, motivation, or search and ranking –, was presented by Trant [2009]. Peters [2009] discusses a large number of studies on the use of folksonomies and particularly on different aspects of retrieval through tags.

### 5.3.2 Web Log Mining

Predominantly, web logs have been used to investigate the query behavior in search engines or the usage of digital libraries in order to better understand such a system's users. This can help webmasters tailor their systems more specifically to the users' needs. A survey on such works about search engines was created by Agosti et al. [2012]. More recently, Thomas [2014] used a combination of controlled user study and web log analysis to identify signals for situations when users were struggling, and found that simple signals, particularly the time spent in a session, are good indicators.

Duarte Torres et al. [2014] found significant differences in the search behavior of young and adult users, for instance, regarding the length of queries or the selection of ranked search results. Examples for the analysis of digital libraries can be found in the works of Nicholas et al. (e.g., Nicholas et al. [2005]). Tagging systems exhibit aspects of both search engines and libraries. While their data is a collection of resources with description and categories, like in a digital library (however not professionally organized), it is created and organized by users in their individual fashion of assigning tags and entering metadata. Nonetheless, the data is clearly more structured than data on the web in general, as posts are constructed according to a specified template.

For the analysis of user behavior in social web systems, request logs have successfully been exploited by Schneider et al. [2009] and Benevenuto et al. [2009]. The gained insights are useful in social studies, they can help improve a system's design and its traffic distribution over the hardware, and they can be used for planning viral marketing and advertisement placement. Benevenuto et al. [2009] collected data from a social network aggregator over a period of twelve days. They found among other things that session durations follow heavy tailed distributions and that users tend to stick with one feature (e.g. photos) within consecutive requests. We conduct a similar analysis in BibSonomy, analyzing the transition probabilities between the retrieval of users, tags, and resources in Section 5.5.3. Schneider et al. [2009] had access to the click streams of large internet service providers and could thus analyze the popularity of individual features in several social networks. They found that the distributions of requests over different features differ from system to system. Similarly, like Benevenuto et al. [2009], they found that especially in the most dominant feature categories (like photos and messaging) users often spend consecutive requests to the same category.

Jiang et al. [2013] presented an analysis of the web logs of the Chinese social network Renren. They look at so-called "latent interactions" (i.e., visits to a page). Among other things, they find that such latent interactions account for the majority of activities in the network and that there are more users who passively consume the content of the network than there are users who actively engage in interaction with others. Further experiments reveal that visits to strangers are rarely reciprocated (even though Renren users can see who visited their content). By and large, latent interactions are "less limited by constraints, such as time and energy, but more meaningful (...) than the social graph" [Jiang et al., 2013]. For social bookmarking systems, such findings raise the question: How strong is the relation between the active contribution of tags and resources and their consumption (in terms of requests to them)? We therefore analyze the popularities in retrieval requests (representing the latent interaction or consumption) and in posts (representing the active contribution) in Section 5.5.4.

### 5.3.3 Web Log Mining in Social Bookmarking Systems

Only very few studies have used web logs in their analysis of tagging systems. Carman et al. [2009] combined tagging data with log data from search engines and compared the distribution of tags to that of query terms in search. They found a large overlap in

the systems' vocabularies, as well as correlations between the frequency distributions of queries and tags to the same URLs. However, they also provide evidence that both tag and query term samples do not come from the same distribution.

While there exists a large amount of literature on tagging systems, to the best of our knowledge, the only work utilizing and analyzing log data from a tagging system are [Millen and Feinberg, 2006, Millen et al., 2007, Damianos et al., 2007]. Millen and Feinberg [2006] investigated user logs of the social tagging system *Dogear* (internally used at IBM) with a focus on social navigation in the system. They found strong evidence that social navigation – users who are regularly looking at bookmark collections of other people – is a fundamental part of the social tagging system. They also found a positive correlation between the assignment frequency of a tag in posts and the frequency of it being used for browsing. These findings have been highly relevant for the understanding of tagging behavior as they provide actual evidence of how users make use of a tagging system's content. Millen et al. [2007] combined log analysis and user interviews to investigate the way users retrieve resources. They observed diverse behavior patterns for different users and found that heavy users tend to spend more time with their own collections than users with only few bookmarks.

Damianos et al. [2007] introduced a tagging prototype called *onomi* to the organization MITRE. They used log data to determine how well the system was accepted and presented several usage statistics from a ten month test period. They found that their users can be categorized into information providers and information consumers depending on their individual ratio of browsing and bookmarking activities.

We compare findings in our experiments to the above mentioned analyses where possible. However, all three works focus on local bookmarking systems located inside the network of a particular company. Therefore, they represent private systems where users only tag resources inside the company's field of interest and hence, the context in which the results were obtained compares only to a certain degree to that of an open, public tagging system. Millen et al. [2007] already note, that company-internal services require their users to use corporate identities instead of pseudonyms, which is typically not the case in public systems. Contrarily, in this work, we focus on the publicly available system BibSonomy to overcome this limitation. This leads to some interesting deviating insights that are discussed in Section 5.5.4 regarding the social and the popularity aspect. While we not only extend the analyses in [Millen and Feinberg, 2006, Millen et al., 2007, Damianos et al., 2007] by investigating a series of aspects of social tagging systems, we also benefit from long-time log data allowing us to get a clearer overview over actual user behavior in an already established social tagging system.

Finally, a recent study by Lorince et al. [2015] analyzed aspects of retrieval in the tagging system last.fm. They did not explicitly use log-file analysis but instead profited from usage information that is made publicly available through the web interface of last.fm. Comparing the use of tagged and untagged content, they conclude that tags increase retrieval only to a minor extent. Since posting resources without tagging them is rarely an option in a tagging system, their analyses are not directly comparable to

ours and rather specific to last.fm. However, where possible we compare our results to theirs and come to similar conclusions.

## 5.4 Dataset

The datasets used in this chapter are based on web server logs and database contents of the social bookmarking system BibSonomy. We restricted the datasets to data that had been created between the start of BibSonomy in 2006 and the end of 2011, since early in 2012 the login mechanism was modified, which introduced significant changes to the logging infrastructure. Anonymized datasets of logs and posts are made available to researchers by the BibSonomy team.<sup>13</sup>

### 5.4.1 User and Content Dataset

We use tagging data from BibSonomy’s database, that is, the folksonomy comprising the users with their posts, containing resources and tags, as well as all data about groups and friendships. In the considered time frame, 852,172 people registered a user account of which 17,932 were classified as non-spammers. They created 551,606 bookmark posts and 2,391,721 publication posts using 250,344 tags.

### 5.4.2 Request Log Dataset

The BibSonomy log files include all HTTP requests (caching is disabled) to the system including common request attributes like IP address, user agent, date, and referer, as well as a session identifier and a cookie containing the name of the logged-in user. Out of the over 2.5 billion requests, we used only those from logged-in non-spammers and additionally filtered out requests to extra resources including CSS, JavaScript, and image files as well as requests from web bots (using a heuristic comparing user agents to those of known bots in various online databases). Furthermore, we removed pages that are irrelevant to our study (like help or administration pages). Additionally, to ensure capturing only actual user behavior, we used a simple heuristic based on the status code of a request’s referer to filter automatic redirects caused by the system instead of by choice of the user (e.g., redirects to the personal user page after editing a post).<sup>14</sup> The remaining dataset contained about 2.4 million requests.

## 5.5 Analysis

In this section, we present our results. For each investigated aspect, we (i) make the research question behind it explicit, (ii) review evidence and arguments related to that

<sup>13</sup><http://www.kde.cs.uni-kassel.de/bibsonomy/dumps/>

<sup>14</sup>This is an improvement over the previously published version of the results in this chapter [Doerfel et al., 2014a] explaining small quantitative (but never qualitative) differences in some results.

aspect from the literature, (iii) present the results of our research and (iv) discuss our findings.

### 5.5.1 The Social Aspect

*With the social aspect, we investigate whether users (re-)use resources that have been shared and tagged by others, either by viewing them or by copying them into their own collection.*

#### Debate in the Literature

The social aspect of tagging has been subject to controversial discussion in the past; and it has been praised and disputed already early in the history of tagging systems. Mathes [2004] stated that folksonomies could “lower the barriers to cooperation” and Weinberger [2005] named it as one of two aspects that “make tagging highly useful”. Marlow et al. [2006] presented an early model for social tagging systems, where they argued that social relations between users are a critical element. The authors pointed out that social interaction connects bookmarking activities of individuals with a rich network of shared tags, resources, and users. Furthermore, Millen and Feinberg [2006] supported the hypothesized social benefits with an analysis of the system Dogear – an internal social tagging service at IBM. They found out that about 74% of all page requests referred to bookmarks that had been contributed by other users. In contrast to that, Damianos et al. [2007] noticed in their system onomi, which also ran in a corporate environment, that users were looking more at their own (70%) than at other users’ collections.

It is not self-evident that similar observations can be made for public bookmarking systems, where users use the service without direct company guidance that might influence their behavior. On the contrary, users may choose to use such systems for individual purposes only, creating their own collections and ignoring the resources of other users. Vander Wal [2005] already pointed out that personal information management may be one of the main reasons why people use social tagging systems, which was also emphasized by Terdiman [2005]. Porter [2005] claimed that “Personal value precedes network value: Selfish use comes before shared use.” A user survey by Heckner et al. [2009] found that about 70% of the interviewed users of bookmarking systems for publications and bookmarks (Connotea and Delicious) claimed to store resources mainly to retrieve them themselves; not particularly to share them. In contrast, for systems to store videos or images (YouTube and Flickr), sharing was the main motivation to contribute. However, it is also noted that “even users of systems who claim that personal information management is very important for them, state that sharing is also part of their motivation of using the systems” [Heckner et al., 2009]. While this survey takes the perspective of *motivation* for posting, we will rather take the viewpoint of the *usage* of posts.

Table 5.1: Visiting content: Request counts in four categories of ownership: requests to the (logged-in user’s) own content, to content from group members or friends, to content from other users, or to general (non-user-specific) pages. Requests to the landing page (see Footnote 15) are excluded from the calculation of the shares.

category	requests	share in %
user’s own	884,525	65.47
groups and friends	44,694	3.31
other users	188,057	13.92
general	233,710	17.30
landing page	296,090	-

## Results

We investigate the different forms of interest in the content of other users through three actions: (i) *Visiting content* is a sign of interest in the material of others. (ii) Similarly, *copying resources* shows a stronger, less casual interest, as it means actively integrating the content into the own collection. (iii) Finally, *copying tags* is an indication that not only the resource was appreciated, but the way it was annotated by another user as well.

**Visiting Content.** First, we analyze the ownership of visited (retrieved) content. We distinguish between four different ownership categories:

- *user’s own*: requests where a user retrieved content (posts) explicitly from their own collection
- *groups and friends*: users retrieved content explicitly from a group they are a member of or from a user they had declared friendship to,
- *other users*: a user retrieved content from a specific other user that was neither a member of any of the user’s groups nor a friend,
- *general*: content was retrieved without specifying a particular user (e.g., a request by tag).<sup>15</sup>

Table 5.1 shows the number of requests (and their shares) in each of these ownership categories. We can observe that roughly two thirds of all requests of logged-in users target their own pages. Users visit other pages in about 35% of the requests to look at either general pages, that is, pages containing posts of several users (about 17%),

<sup>15</sup>The BibSonomy landing page was considered separately, although it lists the most recent posts in the system and thus could be considered a general page. However, many users just visit that page to start their session using the input fields provided on that page and thus ignore the displayed resources.

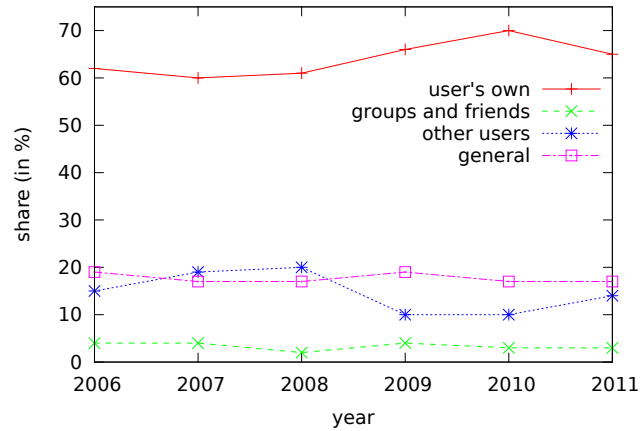


Figure 5.2: Content visits over the years: Per year, shown are the shares of requests to content of different ownership categories.

or content of individual other users or groups (about 17%). Among them, requests to groups and friend pages are both rather infrequent (only about 3%) indicating that these particularly social features (in BibSonomy they are used to control the visibility of posts) play only a minor role. Further, the share of visits to content of others is far below 74%, as reported by Millen and Feinberg [2006] for a company internal tagging system, but similar to the share reported by Damianos et al. [2007]. In summary, we see that the larger share of interactions in BibSonomy happens with the personal collection. However, the interest in other users' content accounts for a significant part – over one third of all retrieval requests – of the interaction with the system.

The previous results are aggregated both over time and over the set of all users. Therefore, next, we examine them first over time and then as distribution over the users. Figure 5.2 shows the shares of requests to content in the four different ownership categories per year. We can observe that requests to content in the own collection (user's own) account for the largest share in every year, roughly between 60% and 70%, dominating the other three categories. The shares of requests to general pages and of requests to groups or friend pages fluctuate only little over the years.

The share of requests to other users exhibits a sharp drop – the share roughly cuts in half – from 2008 to 2009. This drop coincides with slight increases of all other shares, most noticeably that of requests to a user's own content. We can only speculate about possible reasons for that effect: A plausible hypothesis is that, since the system had been available for three years, users had the time to create large collections. Thus such users spend more and more requests on their own collection to retrieve its contents or to navigate through it before they find the resources they were looking for. As this hypothesis concerns the motivation of users, to verify it, one would actually have to ask the users, which is beyond the scope of this investigation.



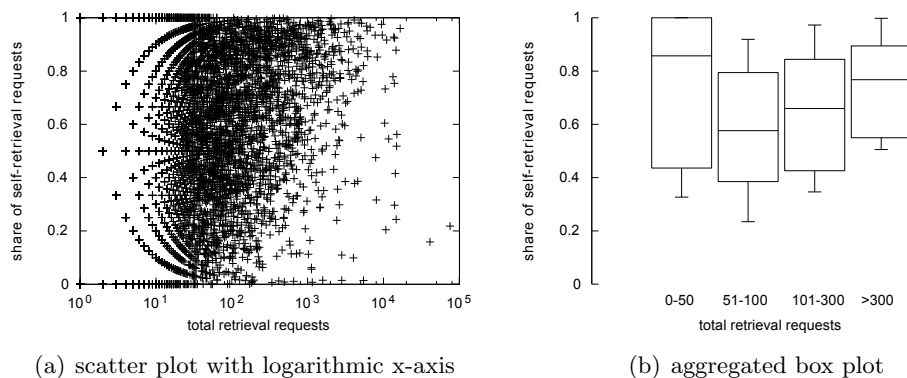


Figure 5.3: Retrieval intensity versus self-retrieval share: Plotted are the total number of retrieval requests versus the share of requests to own content among those requests – in 5.3(a) for each user and in 5.3(b) with users grouped into four buckets by their total number of requests.

However, we can look for further evidence by examining the behavior of individual users: We compare a user’s share of requests to own content to the intensity of this user’s retrieval. Therefore, we determine for each user their total number of retrieval requests (measuring the intensity in which BibSonomy is used by that user to find content), as well as the share of requests to own content among those requests. We can now determine the correlation between these two quantities over the set of all users using Spearman’s rank correlation coefficient  $\rho$  (see Section 2.1.1). We yield  $\rho = -0.42$ , indicating a negative correlation: Users who spend less requests on retrieval in total spend a larger share of those request on self retrieval. The scatter plot of the two quantities, shown in Figure 5.3(a), paints a partially different picture: There are users with about 50 requests or less, who show various individual shares of self-retrieval. However, among the users with more than 50 retrieval requests, we can observe the tendency that with a rising total number of retrieval requests, the focus turns rather on self-retrieval – contrary to the negative correlation that we observed on the full set of users. The plot does not show the number of users each dot represents, but in fact, the set of users with less than 50 requests accounts for about 75 % of all users. Thus, the left part of the plot represents the vast majority of users.

For the box-plots in Figure 5.3(b), we grouped the users by their total number of retrieval requests into four buckets: The first bucket contains all users with less than 50 requests, the second those with 51 to 100 requests, the third those with 101 to 300 requests, and the last bucket all users with more than 300 retrieval requests. Thus, the first bucket represents the majority of users, while the other users are almost equally distributed over the remaining buckets with 8 % in the second and fourth bucket each and 9 % in the third bucket. Although there is a lot of divergence within all four buckets, we can observe two tendencies: Many users with a low total number of

retrieval requests (less than 50) use the system more frequently for self-retrieval than other users. Among the users with more than 50 requests, we observe the tendency that the share of self-retrieval grows with the number of requests. Indeed, if we measure the correlation only among those users with more than 50 requests, we now yield a positive correlation of  $\rho = 0.22$ . These findings strengthen the hypothesis that many users who use the system more than only casually, tend to spend a larger share of their time on their own collections. Therefore, long-time users will be more likely to request their own content than that of other users. Also, the tendency of more active users towards visiting more own than other content, conforms qualitatively with Millen et al. [2007], who similarly observed that a stronger use of the tagging system usually means that more time is spent on the own collection.

**Copying Resources.** When users added new posts to their collections, in 10.7% of all cases, a bookmark or a publication had been copied from another user, as we can see in the first line of Table 5.2.<sup>16</sup> Users copied publications (17.6%) more often than bookmarks (3.5%). We note that the share of 3.5% of copied bookmarks is close to the 2.2% share, reported by Millen and Feinberg [2006] for the IBM-internal system Dogear, while the share for publications (17.6%) exceeds that value by a factor of eight.

One reason for this difference might be the fact that users leave the system when they follow a bookmarked link, while they stay within BibSonomy when they check out details of a publication. Thus, using a bookmarklet (see Section 5.2.1) while visiting the web page to be bookmarked, is the easiest way to post a website. In contrast, for a publication one has already found in BibSonomy, clicking the copy button on its details page is the easiest option. Another factor is that the resource which a user wants to store, must already be available in the system. The second line of Table 5.2 shows that the share of publications that could have been created as copies of an already existing post, is more than twice as high as the respective share of bookmarks.

Taking into account the availability of a resource, the last line of Table 5.2 shows the share of actual copies among possible copies.<sup>17</sup> Of all posts that could have been created through copying at the time of their posting, a share of roughly 40.4% (and even 47.7% for publications) has indeed been copied. This can be regarded as a relatively large share, since looking up publications or websites in BibSonomy is only one out of many possible ways to find interesting bookmarks and publications on the web or elsewhere.

**Copying Tags.** Finally, we study whether not only resources, but also tags are copied. For that purpose, we counted among all post copy operations, how often the copy

---

<sup>16</sup>We ignore imports of bookmark or publication lists (e.g., browser bookmark collections or BibTeX files) because during such transfers of own collections to BibSonomy, it would not be meaningful to look for resources in other users' collections.

<sup>17</sup>The number of posts that have been created as copies, divided by the number of posts where the posted resource had already been available in the system.

Table 5.2: Copying resources: The shares of posts that were and that could have been created as copies of other users' content.

share (in %) of	bookmarks	publications	total
copied posts among all posts	3.5	17.6	10.7
posts that could have been copied	15.6	36.9	26.5
copied posts among posts that could have been copied	22.2	47.7	40.4

was tagged with tags from the user's own vocabulary, and how often tags of the original post were assigned. For 87 % of all copied posts, at least one tag from the own vocabulary was used. In 42 % of all copies, at least one of the original post's tags was adopted. Among the other copy events, 44 % of the original posts had only special tags like "imported", that are probably not meaningful for the user copying the post. Similarly to the copying of resources, we find that users copied tags in a large number of cases; although in the majority of cases (also) own tags were used.

## Discussion

We found evidence for both personal information management and social interaction. In general the findings fit well to the result from Heckner et al. [2009] that the motivation for posting websites and publications is not predominantly social: We found only a relatively low share of visits to groups and friends; and the majority of requests targets content from the logged-in user's own collection. However, while users might not contribute content particularly intending to share it (like in social networks), we could yet observe evidence that they do profit from the availability of other users' content. The shares of visited posts, as well as copied resources and tags, are evidence of social interaction and demonstrate that the collaborative aspect of the bookmarking system is recognized and exploited. For webmasters of such systems our results indicate that it is reasonable to assist users in discovering the content of others, for example, with search functionality or through recommendations.

### 5.5.2 The Retrieval Aspect

*With the retrieval aspect, we refer to the notion that tagging systems are used to manage personal collections of resources for their retrieval later on. We investigate to what degree users make use of their resources and tags after they have stored them.*

#### Debate in the Literature

In a study on the usage of browser bookmarks by Abrams et al. [1998], it was found that users revisit about 96 % of their own bookmarks within one year. Since the

idea of social bookmarking, in a way, is an advancement of the classic browser-based bookmarking, the question arises whether in tagging systems the retrieval behavior is similar to that reported for browser bookmarks. It was hypothesized already at the very beginning of social tagging research that personal information management may be one of the main reasons why people use social tagging systems (e.g., Vander Wal [2005]). Furthermore, as mentioned in the foregoing section, the user survey by Heckner et al. [2009] identified personal management as the main motivation to post web pages or publication references (like in BibSonomy). In the previous section, we saw that the major part of all retrieval requests targets the respective user's personal collection; and such use of own content is part of the personal information management.

For tags, the assumption that they are used to retrieve content later, has been made several times, for example by Vander Wal [2007], by Golder and Huberman [2006], and by Glushko et al. [2008]. A survey by Ames and Naaman [2007] found for several interviewed users that, when annotating photos, self-organization is a primary incentive for using tags. However, they also noticed social aspects to be an important influence for tagging as well. Recently, Lorince et al. [2015] observed for the music tagging system last.fm that using tags only rarely increases retrieval rates – tagged content was listened to about 1.15 times as often as untagged content. This is an indication that tags might not play the presumed role in retrieval.

## Results

We present statistics about revisiting patterns obtained for both publication posts<sup>18</sup> and tags. More precisely, we investigate how many times users revisit *their own* posts and tags and also the time difference between the posting of a resource or tag and its first retrieval, counted in days. In order to give users a reasonable amount of time for revisits, we capture all posts until the end of 2010 and all requests until 2011 (the end of our dataset). This means that each user had at least a whole year to revisit their posted resources and tags; the same time-frame for which Abrams et al. [1998] report at least one visit to almost all bookmarks users kept in their browsers.

The results are shown in Figure 5.4. About 49 % of all publications were revisited by their owner at least once. If a publication has been revisited at all, it mostly was revisited only once (see Figure 5.4(a)). Furthermore, we can observe in Figure 5.4(b) that most of the first revisits to a publication took place shortly after it had been posted, often on the same day. These visits could well be control visits to check the created post, however, it could also mean that users posted a publication immediately before they used it, for example, as a cited reference in a paper they wrote. The revisit figures of tags show a more drastic picture. Only about 17 % of tags are used in queries at least once by a user who previously assigned them to some post. Furthermore, in Figures 5.4(c) and 5.4(d), we can observe similar patterns as for publications:

---

<sup>18</sup>Requests to bookmarks could not be analyzed since they target pages outside BibSonomy and therefore requests for such pages are not recorded in the logs (see Section 5.2.1).

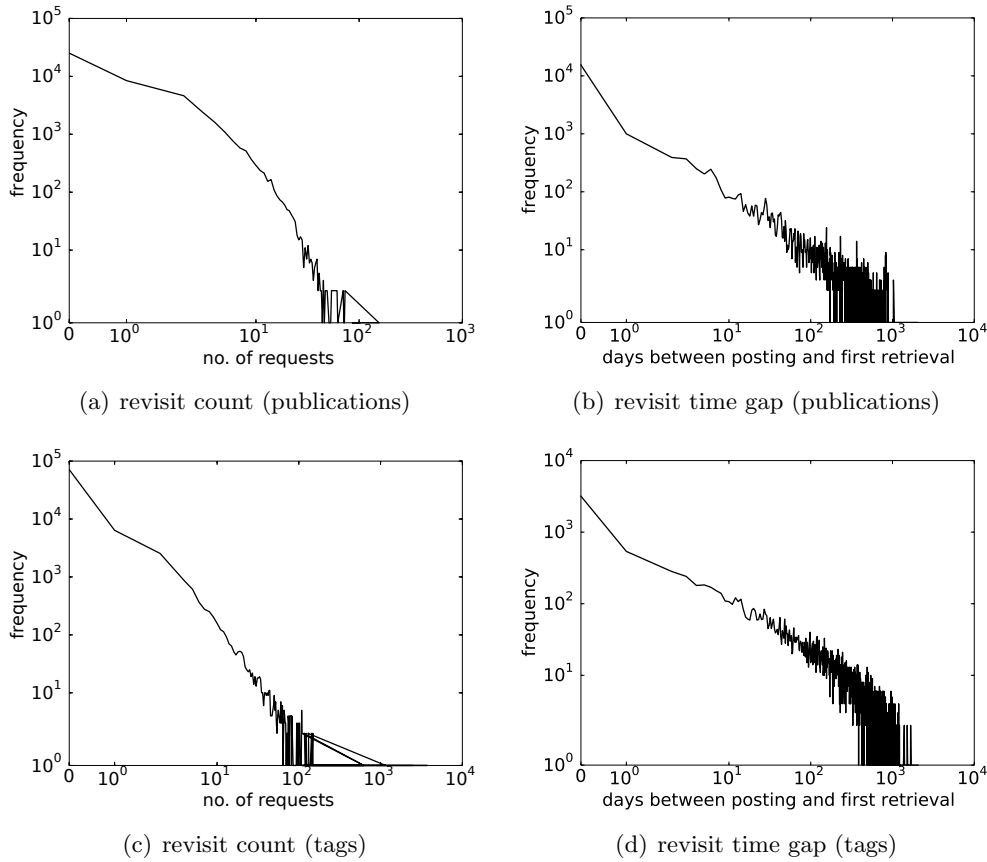


Figure 5.4: Re-visitation behavior of users: Figure (a) illustrates the number of times users revisited their own publications and Figure (b) the number of days elapsed between the posting of a publication and its first retrieval by its owner. Figures (c) and (d) display the revisit count and elapsed days for tags accordingly. All four figures are visualized on a log-log scale.

if revisited, tags mostly only have been revisited once and often shortly after the assignment.

## Discussion

In the previous section, we saw that interactions with the personal collection account for the dominant share of users' retrieval requests. Although according to Heckner et al. [2009], users use the system for later retrieval, we now find that only about half of all publications are revisited. Particularly interesting about this observation is that it does not agree with the work by Abrams et al. [1998] on browser bookmarks, where 96 % of all bookmarked resources got revisited in the time span of one year. The

difference might result from several factors: First of all, using a publication is different to revisiting a website – many websites often renew their content frequently and they usually are easier to consume than scientific publications. Moreover, the user survey reported the difficulty of creating and organizing the bookmarked resources, whereas tagging systems aim to simplify the process of creating and ordering bookmarks as much as possible. This could imply that users tend to store more, simply because the effort is low. Another reason for the lower retrieval rate is certainly that the retrieval of single posts is only one way to make use of the own collection. Another reasonable way of using stored publications (e.g., for citing them) is to mass export them (e.g., simply all or many publications in the collection) into a suitable citation format and selecting the actually used publications offline.

More surprising is the small share of own tags used for retrieval. An explanation for this observation might be that it is reasonable to use many tags for a resource to increase the chance of successful retrieval later on, but for the actual retrieval only a fraction of these tags is sufficient. Furthermore, we will show in the next section that using tags is not the dominant way to query BibSonomy. Finally, we note that our results are in line with the conclusions of Lorince et al. [2015], who found that in last.fm, tagged content is only slightly more often retrieved than untagged content. Due to the fact that BibSonomy (like most other tagging systems) does not support posting without tagging, these results are not directly comparable. Still, both findings are evidence that the role of tags in retrieval is not as strong as has been previously assumed.

For webmasters of a tagging system, our results indicate that visits of resources or tags, could be an important measure (distinguishing between visited and unvisited content) and should be considered in the assessment of the importance of resources or tags, for example in such features as ranking of search results or tag recommendations. Until now, the dedicated algorithms typically focus rather on the number of posts that a resource or tag occurs in (e.g., those mentioned here in Sections 2.4.3 and 7.3) than on their retrieval.

### 5.5.3 The Equality Aspect

*With the equality aspect we focus on the question whether the three entity sets in a tagging system – the sets of users, tags, and resources – are equally important for navigation or retrieval.*

#### Debate in the Literature

A folksonomy – the structure underlying tagging systems – has been defined as a quadruple  $\mathbb{F} = (U, T, R, Y)$  consisting of the sets of users  $U$ , resources  $R$ , and tags  $T$  together with the tag-assignment relation  $Y \subseteq U \times R \times T$  (see Section 2.3.1). In that model, users, tags, and resources are treated equally and in fact even symmetrically. The folksonomy model has been widely accepted and many algorithms build on it, for

example, the *FolkRank* (Section 8.2) or the tensor factorization method by Rendle et al. [2009]. Since tag assignments link entities of all three sets together, the idea of the typical folksonomy navigation is that these entities can be reached following these links back and forth (e.g., clicking a tag to request all posts to which that tag has been assigned). A counterargument to the symmetry of tags, resources, and users is the fact that tag assignments usually occur in groups, which are represented by the posts of a tagging system: Each post is created by *one user*, who assigns *several tags* to *one resource*. Thus, one post usually provides links to the one user and to the one resource, but to more than one tag. Furthermore, although it is typically assumed that tags are added to posts as a means to retrieve the posted resource, in the previous section, we saw evidence that tags might play a less dominant role than expected in a *tagging* system.

## Results

We discuss this aspect in two parts: First, we analyze the shares among all requests with respect to the entities they target (either users, tags, or resources). Then, we investigate the probability of transitions between entities of different classes in the users' navigational paths.

**Request shares.** Like for the previous aspects, we analyze retrieval requests. We split them into requests querying specifically for users, tags, or resources.<sup>19</sup> Requests with more than one queried entity have been assigned to the set of that entity that dominates the request. For example, a post's details page belongs to the post's owner, but the target is clearly the resource rather than the user. A request containing a user and a tag has been counted as a tag request. Requests that are not specific to some entity (like the landing page) have been ignored.

For each class of entities (users, tags, and resources), the first row of Table 5.3 shows the average number of requests per entity. The second and third row show the total number of requests to entities of one class and their respective relative shares among the total number of retrieval requests to any of the entities. For comparison, the other lines of the table similarly report the requests to entities and their shares, however either considering only retrieval of content from the own collection (to self) in lines four and five, or, contrarily, considering only requests where users have accessed content of other users (to other) in lines six and seven. We can clearly see that the total request numbers are not equally distributed: There are about 2.1 times more requests to specific users than to specific tags. The share of resources is slightly higher than that of tags. From the average number of requests per entity, we can deduce that this strong imbalance is not caused simply by a similar imbalance in the size of the

<sup>19</sup>Note that requests to resources are generally underrepresented due to the lack of recorded requests to bookmarks (see Section 5.2.1).

<sup>20</sup>For tags, for example, that is dividing the total number of requests to any tag by the total number of tags in BibSonomy.

Table 5.3: Entity request shares in BibSonomy: We report for each class of (folksonomic) entities (users, tags, and resources) the average number of requests per entity in that class,<sup>20</sup> as well as the total number and relative share of requests to entities of that set – among all requests (total), among requests to the user’s own collections (to self), and among requests targeting content outside the own collection (to others).

requests	user	tag	resource
per entity	30.33	1.08	0.14
total	543,837	269,212	316,582
total (in %)	48.14	23.83	28.03
to self	435,513	192,737	217,587
to self (in %)	51.49	22.79	25.72
to other	108,324	76,475	98,995
to other (in %)	38.17	26.95	34.88

three classes. Despite the fact that BibSonomy has far more tags and resources than it has users, on average a user page is queried much more often than a resource or a tag page.

The use of a tagging system consists of both working with one’s own collection as well as working with posts from other users; and we already found in Section 5.5.1 that requests to the own collection dominate those to content of other users. From the figures in Table 5.3, we can observe that for requests to the own collection (to self), the share of requests by user increases slightly compared to the share among all requests (total). This is not surprising, as all requests to the own collection must necessarily be requests to a user: to oneself. Thus, among these, tag or resource requests are those that have two targets (a user and a tag or resource) and have been classified into one of either category by the rule mentioned above. Looking at the requests to content outside the own collection (to other), we observe that the share of user requests drops, compared to the full set of requests. Nevertheless, the queries for users still outnumber those for tags, however to a lesser extent. It is also interesting to note that the ratio between requests to tags and to resources is roughly comparable: 1.2 (total requests) versus 1.1 (to self) versus 1.3 (to others). This indicates a comparable user behavior within one’s own collection and within the content of other users.

With the above mentioned assignment of each request to one dominating entity, we chose a rather conservative approach that tends to *underestimate* the relevance of requests to users. As mentioned above, we counted each request with multiple queried entities only once, for the dominant entity in that request, which in all cases was always a tag or a resource. Therefore, in a similar experiment, we directly counted the requested entities. For example, a request with requested resource and requested user was counted for both user and resource. The results are qualitatively comparable and



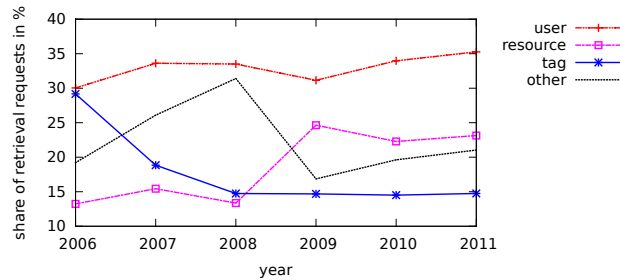


Figure 5.5: Request shares over time: Plotted are the shares of retrieval requests each year to the different classes of entities: users, tags, and resources. All remaining retrieval requests that target neither type of entity, are summarized as ‘other’.

show an even stronger imbalance towards users (about 63% of the requested entities were users).

The figures in Table 5.3 indicate that users are the main means of navigation in BibSonomy, rather than tags as one might have expected in a *tagging* system. To gain deeper insights, in the following, we look at similar figures over time. We investigate the full set of retrieval requests, that is, next to requests to users, resources, or tags, we also count all other retrieval requests (e.g., requests to the full-text search,<sup>21</sup> requests by author or by a publication’s  $\text{BIB}_{\text{T}}\text{E}_{\text{X}}$  key, or requests to pages listing the recently most popular resources or tags). Figure 5.5<sup>22</sup> shows how these shares develop over time. We can observe that in the first year, the numbers of requests to tags and to users were almost equal. The share of requests to users stayed roughly the same over the years. However, within two years, the (relative) share of requests to tags drops significantly and stays almost constant afterwards. At the same time, the share of other requests increases. This indicates that users have found other means of navigation rather than using tags for retrieval. Since BibSonomy has constantly been extended with new features, it is natural, that users would use these new features (e.g., a full-text search) and therefore others, like tags, to a lesser extent. Rather surprising is, however, the fact that only the share of requests to tags shrinks.

One feature that is particularly suitable to retrieve resources is the full-text search. As it is not part of the folksonomy structure, which underlies a tagging system (see above), we have omitted it in the previous analyses. However, since we saw that the role of tagging is not as dominant as one would expect, in the next analysis, we compare the requests using tags to those using the full-text search. Table 5.4 shows, similarly to Table 5.3, the absolute number of requests together with the shares (among requests to either search or tags). Again, we distinguish between such requests in general (total), requests to the own collection (to self), and requests to content of

<sup>21</sup>The full-text of a publication in BibSonomy is its collection of metadata.

<sup>22</sup>The shares in Table 5.3 are computed using only of those retrieval requests that target either users, resources, or tags. Through the inclusion of the “other” requests, the percentages in Figure 5.5 are not directly comparable to those in the table.

Table 5.4: Tag-based retrieval versus search in BibSonomy: Reported are the total number and relative share of requests to either tag pages or to search – among all requests (total), among requests targeting own content (to self), and among requests targeting content outside the own collection (to others), for example, requests to user Y by user X.

requests	tag	search
total	269,212	141,843
total (in %)	65.49	34.51
to self	192,737	20,663
to self (in %)	90.32	9.68
to other	76,475	121,180
to other (in %)	38.69	61.31

others (to other). We can observe that overall, requests using tags outnumber those using the search roughly two to one. However, the choice of either means depends clearly on the scope of the search: If the own collection is targeted, requests with tags are used roughly nine times more often than the full-text search. On the other hand, requests to content outside the own collection are more often conducted with the full-text search (about 1.6 as many requests as with tags). These figures indicate that users indeed make use of tags, yet rather when they retrieve their own resources; after all it is “their” tags they used to annotate them with in the first place. For content they did not annotate themselves, the full-text search is more often the preferred option.

**Transition Probabilities.** Next, we look at navigational transition probabilities between the entity classes users, resources, and tags (e.g., the probability of requesting a tag after requesting a user). We determine the transitions using each requests’ HTML referer attribute and compute first order Markov chain probabilities from one entity set to another. Figure 5.6 shows the results as a graph. We can observe that self-transitions account for the largest shares of requests starting from a tag or a resource page. This suggests that often, users tend to stay with the same type of (requested) entity in their navigational paths through BibSonomy. This observed share of self-transitions is consistent with findings in social networks by Schneider et al. [2009], who also observed that users tend to stick with the same feature in consecutive requests.

Requests from a user page are distributed almost equally between requests targeting a user or a resource, while a slightly smaller share falls upon tag pages. Aside from that, there are a lot of transitions from user pages to resource pages and tag pages. This is not surprising, as user pages consist of listings of a user’s resources and their tag clouds, such that both can be reached with a single click. It also explains the

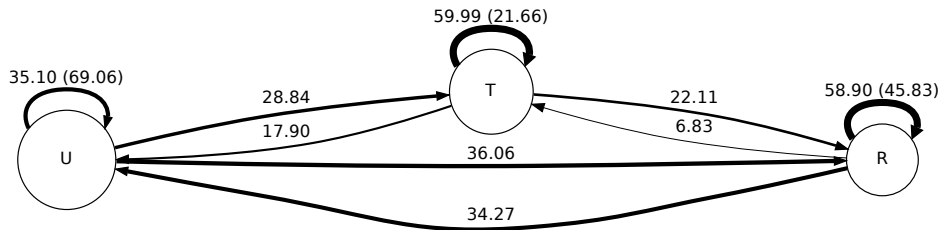


Figure 5.6: Transition probabilities between the three classes of entities in BibSonomy: The nodes **U**ser, **T**ag, and **R**esource correspond to the columns in Table 5.3, and their sizes reflect the total number of requests to entities of these sets. The edges represent transition probabilities from a page of one entity set to another. The percentage in brackets on the self-loop edges describe the fraction of pagination or reload effects on navigation.

relatively frequent transitions back to user pages (particularly from resource pages) and symbolizes the “browsing” in the system. Only few requests are transitions from a resource page to a tag page. This means that only rarely, users seem interested in resources with the same tags as the resource at hand.

We also looked at the fraction of page self-transitions, where the referer and actual URL of a request are the same. This effect occurs typically from page reloading or pagination effects, such as viewing the next 20 elements of a longer list of publications. In Figure 5.6, these fractions (of the entity class self-transitions) are shown as percentages in brackets. For example, about 69% of the requests that lead from a user page to a user page, actually lead from a user page to itself again. This might be explained by the fact that most users’ collections exceed the amount of items which are displayed at the same time, so users have to “turn the page” to view the next items. Interestingly, this effect is greatly diminished on tag pages, where only 22% of the transitions from tag to tag are actually self-transitions. This could mean that many transitions lead from a tag page to a more refined tag page (e.g., by selecting an additional tag in the following request). The relatively high amount of self-transitions on resource pages (about 46%) mainly stems from exporting the resource into a particular format.

## Discussion

We have observed a strong *inequality* between the use of the three folksonomy entities of users, tags, and resources. While the numbers of requests to tags and to individual resources are similar, they are dominated by the requests to user pages. This is surprising, as there are fewer user pages than tag or resource pages available in BibSonomy. When discussing navigation within folksonomies, resources are usually regarded as targets of queries. As navigational means to find or retrieve these resources, often tags – rather than users – receive the larger interest, as they can function as resource descriptors. In BibSonomy, it seems, however, that the user pages

are the main means of navigation and tags are mainly used to navigate through ones own collection rather than through the system. From the transition probabilities, we saw that especially navigation from resources to tags (and thus to potential further resources to the same tag/topic) is rather rare. The unexpected observation that tags do not play the main role in the users' navigation behavior, has consequences for webmasters who run and design such systems: Algorithms, like *FolkRank*, that model the transitions between entities, need to be revisited. There, transitions between users, tags, and resources are modeled with equal probabilities, which – as we found out – does not properly reflect actual user behavior.

### 5.5.4 The Popularity Aspect

*The popularity assumption concerns the practice of measuring an entity's popularity by counting the number of posts it occurs in. We investigate to what extent this popularity of folksonomic entities – the number of posts a user, a resource, or a tag occurs in, or its frequency distributions – matches similar properties in requests.*

#### Debate in the Literature

In tagging systems, the notion of popularity is usually exploited in several ways: (i) special “popular” pages summarize the most frequently posted resources or tags, (ii) next to a resource, the number of posts it occurs in is shown, (iii) users' profile pages often show the number of their posts, and (iv) several algorithms for the recommendation of tags [Jäschke et al., 2008] and resources [Bogers, 2009] suggest the most frequently used entities. Perhaps the most prominent application of tag frequencies are tag clouds, where the frequency of a tag corresponds to its font size and particularly rare tags sometimes are not displayed at all. Brooks and Montanez [2006] point out that it is taken for granted that the tags a user assigns, are the same as those a reader would select. Hence, the authors identified the relationship between the task of article tagging and information retrieval as an open question to investigate. In the user study by Sinclair and Cardew-Hall [2008], it was found that tag clouds are perceived as visual summaries of resources, and that clicking into tag clouds requires less cognitive effort than entering search queries. This indicates that the size of a tag is indeed relevant for the users in their query behavior, but to the best of our knowledge, the correlation between tag usage in posts and requests has not yet been investigated in a large-scale scenario in a public tagging system. For the company internal system Dogear, Millen and Feinberg [2006] reported a correlation of 0.67 between the frequencies of a tag in posts and in requests. Contrary to the often assumed connection between tags' popularity in posts and their importance in retrieval, Lorince et al. [2015] found that it is rather the more idiosyncratic, less often used tags that lead to higher retrieval rates.

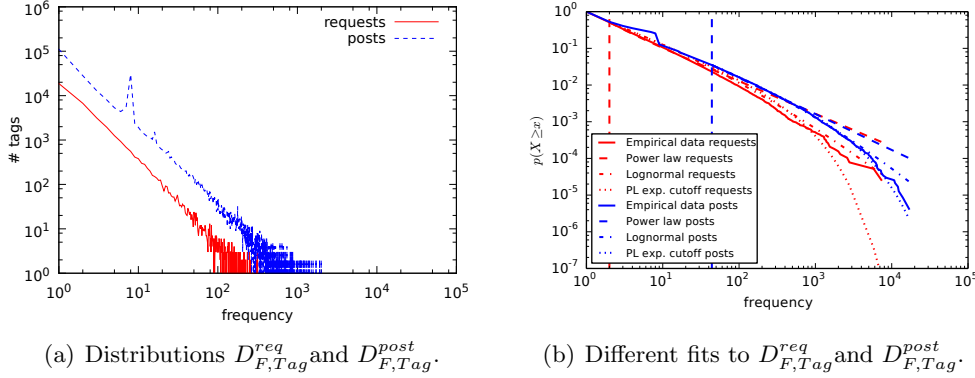


Figure 5.7: Frequency distributions of tags in requests and posts: In log-log scale, displayed are (a) the frequency distributions for tags in requests ( $D_{F,Tag}^{req}$ ) and for tags in posts ( $D_{F,Tag}^{post}$ ), and (b) fits of the respective complementary cumulative probability distributions to different standard cumulative probability distributions (the vertical lines indicate the corresponding  $x_{min}$  values).

Regarding the overall behavior, Cattuto et al. [2007] noted that frequencies of entities in posts follow heavy-tailed distributions (mostly clean power-law fits), and thus could be the result of preferential attachment.

## Results

Since tag clouds are one of the most common application of popularity in social bookmarking systems, we begin the investigation of the popularity aspect by looking at tags. Afterwards, we analyze the same questions for users and resources.

**Tags.** We start the discussion of popularity of tags by analyzing their distributions of frequencies in the request logs ( $D_{F,Tag}^{req}$ ) and in the posts ( $D_{F,Tag}^{post}$ ).<sup>23</sup> More precisely:

- $D_{F,Tag}^{req}(k)$  counts how many tags have been *requested*<sup>24</sup> exactly  $k$  times. Thus,  $n = D_{F,Tag}^{req}(k)$  means that exactly  $n$  tags have been requested exactly  $k$  times.
- $D_{F,Tag}^{post}(k)$  counts how many tags have been *assigned* to exactly  $k$  posts, and thus constitutes the usual node degree distribution described in [Cattuto et al., 2007].

<sup>23</sup>We ignore posts from two users who are known to create posts solely automatically from publication catalogs to provide more content in the system.

<sup>24</sup>In the request distribution, we do not distinguish between requests made by clicking on tags (e.g., in a tag cloud or next to a post) or by entering them directly into the tag search field since these types of requests are indiscernible in the logs.

Both distributions are shown in Figure 5.7(a).<sup>25</sup> The first observation is, that  $D_{F,Tag}^{post}$  dominates  $D_{F,Tag}^{req}$ , meaning that in total there are more tag assignments than requests for tags. Since tag frequency distributions in posts ( $D_{F,Tag}^{post}$ ) are known to be heavy-tailed [Cattuto et al., 2007] – mostly power law – we expect that the distribution of tag frequencies in the request logs ( $D_{F,Tag}^{req}$ ) has similar properties. To confirm this, we fitted the power-law function ( $y = cx^{-\alpha}$ , for  $x > x_{min}$ ) to the empirical data as described in Section 2.1.1, and we compared the resulting fit to the exponential function as a lower barrier for heavy-tailed distributions, as well as the lognormal function and the power-law function with an exponential cutoff (which means that for large  $x$  values the function deviates from the typical power-law function). We visualize the empirical distributions, the best power-law  $x_{min}$  values (vertical lines), and the corresponding fits in Figure 5.7(b) for both  $D_{F,Tag}^{post}$  as well as  $D_{F,Tag}^{req}$ .<sup>26</sup> For the fits of the power-law function, we obtained  $\alpha = 1.98$  and  $x_{min} = 44$  for  $D_{F,Tag}^{post}$ , and  $\alpha = 1.89$  and  $x_{min} = 2$  for  $D_{F,Tag}^{req}$ . The distributions are similar with regard to their slopes  $\alpha$ . Noteworthy is the higher result of  $x_{min}$  for  $D_{F,Tag}^{post}$  (in contrast to the small value for  $D_{F,Tag}^{req}$ ), indicating that the power-law fit only holds for a smaller portion of the distribution (the tail). Visual inspection suggests that there are slightly fewer tags with low frequencies than one would expect in a power-law distribution. While an in-depth analysis of this phenomenon is beyond the scope of this work, we can speculate that it might be a consequence of the use of tag recommenders that typically suggest tags that are already frequently used, leading to an ignorance of low frequency tags.

A comparison between the fits to the other candidate distributions showed that the power-law function is a statistically significantly better fit to the data than the exponential fit. Both the lognormal as well as the power-law function with an exponential cutoff are also good fits to the data, confirming our assumption about heavy-tailed distributions. Also, they are even slightly better fits to the data than the pure power-law function, as can be seen in Figure 5.7(b): Both distributions exhibit a slight decay – visible where the line of the empirical distribution ( $D_{F,Tag}^{req}$  at  $\approx 10^2$  and  $D_{F,Tag}^{post}$  at  $\approx 10^3$ ) falls below the straight line of the respective power-law fit. We hypothesize that effects of information filtering might be a factor in this deviation from power law, influencing the preferential attachment process in the way Mossa et al. [2002] showed. Similar arguments were made by Cha et al. [2009] for popularity distributions in video portals. Nevertheless, detailed investigations regarding this cutoff are necessary for a better understanding of this behavior. By and large, the distributions suggest similar processes of how users post tags and how they request them (i.e., processes yielding heavy-tailed distributions).

<sup>25</sup> A close investigation of the notable peak in the distribution  $D_{F,Tag}^{post}$  at frequency 8 reveals that this anomaly is due to the activities of one single user, who used 28,989 tags exactly 8 times. We therefore ignore the peak in the following discussion.

<sup>26</sup> For better readability we omitted the (weak) exponential fit.

Table 5.5: Pearson’s correlation coefficient  $r$ , Spearman’s rank correlation coefficient  $\rho$ , and the Jensen-Shannon divergence  $JS_2$  for pairs of distributions. In each row, a distribution  $D_{Entity}^{req}$  (Entity is either Tag, User, or Publication (Pub)) of requests (or their frequencies  $D_{F,Entity}^{req}$ ) is compared to a distribution  $D_{Entity}^{post}$  of posts (or their frequencies  $D_{F,Entity}^{post}$ ).

requests	posts	$r$	$\rho$	$JS_2$
$D_{F,Tag}^{req}$	$D_{F,Tag}^{post}$	0.968	0.596	0.052
$D_{Tag}^{req}$	$D_{Tag}^{post}$	0.420	0.059	0.440
$\emptyset D_{Tag}^{req}$	$\emptyset D_{Tag}^{post}$	0.414	0.517	0.271
$D_{F,User}^{req}$	$D_{F,User}^{post}$	0.942	0.242	0.197
$D_{User}^{req}$	$D_{User}^{post}$	0.092	0.548	0.492
$\emptyset D_{User}^{req}$	$\emptyset D_{User}^{post}$	0.081	0.712	0.471
$D_{F,Pub}^{req}$	$D_{F,Pub}^{post}$	0.823	0.803	0.329
$D_{Pub}^{req}$	$D_{Pub}^{post}$	0.554	0.032	0.707
$\emptyset D_{Pub}^{req}$	$\emptyset D_{Pub}^{post}$	0.609	0.252	0.152

Further, we directly compare both  $D_{F,Tag}^{post}$  and  $D_{F,Tag}^{req}$  with each other using Pearson’s correlation coefficient  $r$  and Spearman’s  $\rho$  (see Section 2.1.1).<sup>27</sup> From the first row in Table 5.5 we can observe that the Pearson’s and Spearman’s correlations are high. An explanation for the smaller Spearman’s  $\rho$  value is the fluctuation in the distributions (see Figure 5.7(a)) where the number of tags no longer decreases monotonously with increasing frequency. Finally, a comparison of the distributions using the Jensen-Shannon divergence  $JS_2$ <sup>28</sup> confirms similarity.

<sup>27</sup>Note that all correlation results in this section are statistically significant with a p-value below 0.05, which is why we do not report it explicitly for each calculation.

<sup>28</sup>The Jensen-Shannon divergence  $JS_b(P||Q)$  [Lin, 1991] is a measure for the similarity of two probability distributions  $P$  and  $Q$ , computed as  $JS_b(P||Q) := \frac{1}{2}KL_b(P||M) + \frac{1}{2}KL_b(Q||M)$ , with  $M := \frac{1}{2}(P + Q)$ . It is a derivative of the Kullback-Leibler divergence [Kullback and Leibler, 1951]:

$$KL_b(P||Q) := \sum_{x \in X} P(x) \log_b \frac{P(x)}{Q(x)}.$$

In contrast to  $KL_b$ ,  $JS_b$  is symmetric and well-defined for any two probability distributions –  $KL_b(P||Q)$  is undefined when there is some  $x \in X$  with  $P(x) > 0$  and  $Q(x) = 0$ . The latter is relevant in our study as there are tags that are used in posts but not in requests and vice versa. The value of  $JS_b$  is always non-negative and depends on the choice of the logarithm base  $b$ , however, only by a constant factor. The choice of the dual logarithm results in an upper bound of 1 for  $JS_b$ .

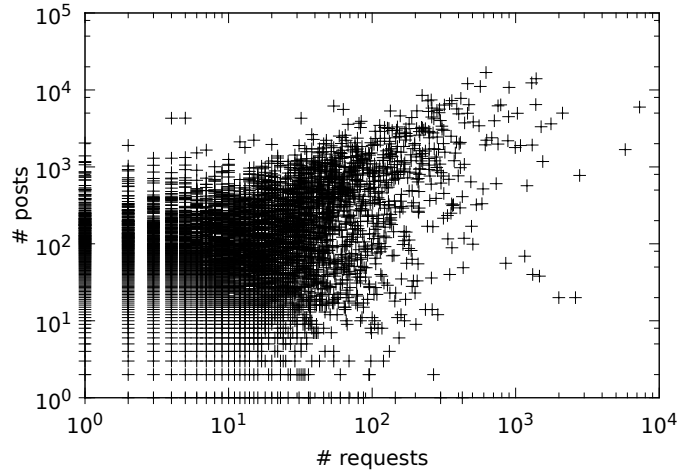


Figure 5.8: The scatter plot (in log-log scale) of the numbers of requests to a tag  $t$  versus the number of post a tag  $t$  occurs in. Only for higher frequencies, the number of requests and posts of a tag appear to be correlated.

In the tag frequency distributions, we found similarity in *the ways* users assign and request tags. As a next step, we analyze the tag popularity on the level of individual tags, to see whether there are similarities regarding *which* tags users assign and request. Particularly, we look at the distributions  $D_{Tag}^{req}$  and  $D_{Tag}^{post}$ , where

- $D_{Tag}^{req}(t)$  is the number of *requests* to a tag  $t$  (i.e.,  $n = D_{Tag}^{req}(t)$  means that the tag  $t$  has been requested exactly  $n$  times) and
- $D_{Tag}^{post}(t)$  is the number of *posts* that the tag  $t$  occurs in.

Figure 5.8 shows the scatter plot of these two tag distributions, where each point in the diagram denotes one tag  $t$  with its number of requests  $D_{Tag}^{req}(t)$  and its number of posts  $D_{Tag}^{post}(t)$  as coordinates. We can immediately see that despite the similarity in the behavior of tag frequencies, there are enormous differences on the level of individual tags. Only for very frequent tags (more than 100 requests) one could presume a correlation between both frequency counts. To quantify the effect, the second row of Table 5.5 shows the correlation coefficients and the Jensen-Shannon divergence for the two distributions  $D_{Tag}^{req}(t)$  and  $D_{Tag}^{post}(t)$ . We can observe rather low correlation and a much higher divergence. This means – contrary to the popularity assumption – that the number of posts a tag is assigned to, and the number of times a tag is queried, are only mildly correlated. The found correlation of  $r = 0.42$  is also lower than the one reported for the company system Dogear (0.67).

A closer look at the log data revealed that many tags which have been used in posts were never queried at all, and several tags have been queried but were never assigned to any post. Therefore, we look at similar distributions as before but we



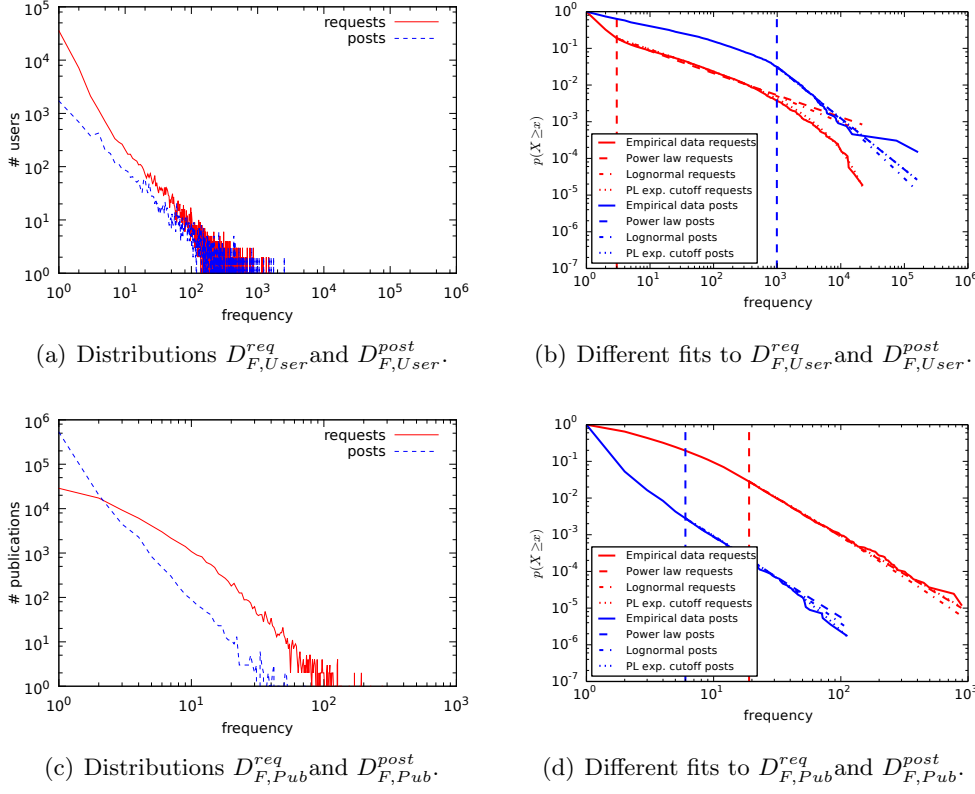


Figure 5.9: Frequency distributions of users and publications in requests and posts: In log-log scale, displayed are (a) the frequency distributions for users in requests ( $D_{F,U_{ser}}^{req}$ ) and for users in posts ( $D_{F,U_{ser}}^{post}$ ), and (b) fits of the respective complementary cumulative probability distributions to different standard cumulative probability distributions (the vertical lines indicate the corresponding  $x_{min}$  values). Accordingly, Figure (c) shows the frequency distributions of resources (publications) ( $D_{F,P_{ub}}^{req}$  and  $D_{F,P_{ub}}^{post}$ ) and Figure (d) the corresponding fits.

specifically ignore tags that only occur in one of the two tag distributions. We yield distributions  $\emptyset D_{Tag}^{req}$  and  $\emptyset D_{Tag}^{post}$ , reducing the number of considered tags significantly to only 11%. Their distributions' correlations and divergence can be found in the third row of Table 5.5. We can observe that the limitation to such “active” tags yields strong Spearman correlations and less divergence, as the active tags' rankings exhibit far less ties than the full set of tags.

**Users and Publications.** As with tags, we investigate similar distributions of both users and resources:  $D_{U_{ser}}^{req}$  counts the requests to specific users,  $D_{U_{ser}}^{post}$  counts a user's posts,  $D_{P_{ub}}^{req}$  counts the requests to a particular publication, and  $D_{P_{ub}}^{post}$  counts the posts

containing a publication. Similarly, we have the according frequency distributions (e.g.,  $D_{F,User}^{req}$ ) and the restricted distributions to active entities ignoring those that occur either only in posts or only in requests (e.g.,  ${}^{\emptyset}D_{Pub}^{req}$ ). Again, we restrict resources to publications (and thus omit bookmarks), as visits of bookmarks are not recorded in the log files (see Section 5.4). The correlation results are depicted in rows four through nine in Table 5.5 and the frequency distributions are illustrated in Figures 5.9(a) through 5.9(d).

The distributions of user (publication) frequencies in requests  $D_{F,User}^{req}$  ( $D_{F,Pub}^{req}$ ) and in posts  $D_{F,User}^{post}$  ( $D_{F,Pub}^{post}$ ) are similar and yield relatively high correlation according to Pearson's  $r$  (Table 5.5, rows four and seven). Their Jensen-Shannon divergences  $JS_2$  are higher than for tags, but still the distributions are relatively similar. Since the distributions  $D_{F,Pub}^{post}$  and  $D_{F,Pub}^{req}$  are for the most part monotonically decreasing (Figure 5.9(c)), their rank correlation is high, unlike for the frequencies of users (Figure 5.9(a)). Notable in both cases (users and publications) is that the distributions of frequencies in posts and requests are no longer “parallel” as they were in the case of tags (compare Figure 5.7(a) with 5.9(a) and 5.9(c)). Power-law fits for the publication frequency distributions of both posts  $D_{F,Pub}^{post}$  ( $\alpha = 3.17$ ,  $x_{min} = 5$ ) and requests  $D_{F,Pub}^{req}$  ( $\alpha = 3.04$ ,  $x_{min} = 22$ ) are decent fits with relatively low  $x_{min}$  values (see Figure 5.9(d)). Not surprisingly, the fits of the power-law function are statistically significantly better than those of the exponential function. However, it is extremely difficult to distinguish the fits of the lognormal function and the power-law function with exponential cutoff from the power-law fit – a strong indicator for the presence of heavy-tailed distributions. For user frequencies, our results also indicate a good power-law fit for  $D_{F,User}^{req}$  ( $\alpha = 1.60$ ,  $x_{min} = 3$ ). For  $D_{F,User}^{post}$ , we yield  $\alpha = 2.39$ ,  $x_{min} = 988$  and thus a fit for only a small part of the distribution. The fits are shown in Figure 5.9(b). Similarly to our investigations on tag frequencies, we obtain a higher  $x_{min}$  value for the frequencies in posts than for those in requests, although this time, the  $x_{min}$  value is much higher. For  $D_{F,User}^{req}$  all candidate functions are better fits than the exponential function; both the lognormal as well as the power-law function with exponential cutoff are better fits to the data than the pure power-law function. The power law with cutoff is even better than the lognormal. For  $D_{F,User}^{post}$  the power-law fit is better than the exponential function and it is difficult to distinguish from the other candidate distributions.

Regarding individual entities, we again measure correlations between the respective distributions in Table 5.5 (for users in rows five and six and for publications in rows eight and nine): For resources ( $D_{Pub}^{req}$  and  $D_{Pub}^{post}$ ), we obtain similar results as previously for tags: Pearson's correlation is moderate, the divergence is even higher than for tags, there is almost no rank correlation, and removing “inactive” publications (occurring either only in posts or in requests) yields higher rank correlation and lower divergence. The elimination of such publications leaves only about 12% of the original set of publications. By and large, we find only moderate correlation even among the actively posted and requested publications. A possible explanation for the correlation

results might be based on the large number of publications that only get posted and requested infrequently. Slight changes in the post or request counts (e.g., once instead of twice) only change Pearson’s correlation slightly, but have a large influence on Spearman’s correlation. For users ( $D_{User}^{req}$  and  $D_{User}^{post}$ ) we find different behavior: almost no correlation according to Pearson’s  $r$  and moderate to strong rank correlations  $\rho$  (higher than for tags and publications). This indicates that users with many posts indeed tend to be requested more, but not proportionally more.

### Discussion

The obtained results do not clearly support the initial assumption. The overall behavior of tag (and to a smaller degree of user and resource) frequencies is similar in requests and posts and they are heavy-tailed as expected. In almost all examples, we can find a good power-law fit. However, in some occasions the distribution decays from the straight power-law function, which indicates the presence of other heavy-tailed distributions. This behavior might be based on distinct processes creating these distributions, warranting further detailed investigations in the future.

On the level of individual entities, we observe weaker correlations and only among the more actively used entities. It is surprising that, despite the fact that tag clouds are displayed in BibSonomy and users can click tags to find according resources, the choice of tags in requests is not stronger correlated to their popularity in posts. Also, we noted a strong difference to the company internal system Dogear, where much stronger correlations could be observed for tags. For operators of a tagging system, the results indicate that it is reasonable to exclude rarely requested tags completely from tag clouds or to use request frequencies instead or in addition to post frequencies in tag clouds. These could even be personalized to a user’s query behavior.

## 5.6 Conclusion

In this chapter, we tackled a number of prominent research questions about social tagging systems using a web server log dataset from the scholarly system BibSonomy containing data on both posts and requests. We have thus supplemented previous work – that has tapped into surveys and post data – by also reflecting *actual user behavior* leveraging request data. Our findings paint a rather mixed picture about the four aspects we studied:

(RQ1) *The Social Aspect*: In our analysis of the social tagging system BibSonomy, we found evidence both for and against the assumption that the activities in a tagging system are primarily *social*. While some user actions indeed indicate social sharing, others are evidence for individual purposes. Furthermore, we could observe that resources are reused by others, especially publications are copied often. This suggests that both kinds of activity are relevant in a social tagging system and should therefore be supported in the system’s design. Also, it is encouraging news for tagging in general, as it fits to the idea that users contribute to the system for their own purposes, but

they can still profit from the contributions of others, which justifies the collaborative nature of social bookmarking.

We saw that users who just started with the system tend to spend more requests on the retrieval of own content. Given time, users start showing interest in the resources of others, however, with rising number of requests, the share of requests to the own collection rises. A possible interpretation of this result could be that long time users need more support to navigate their own collections. One such approach could be recommenders for own content (traditionally, folksonomic recommendations have focused on recommending new resources), or clean-ups, removing resources that have never been revisited or that have not been used in a long time, to reduce the size of one's collection and thus make self-retrieval more efficient.

Finally, we saw that explicit social ties, like groups and friendship relations, play only a minor role for retrieval in BibSonomy. One reason might be that the main advantages of these features are related to BibSonomy's visibility concept. To encourage users to network, more features exploiting such ties could be helpful. In BibSonomy (meanwhile), users are assisted during the posting process. When the resource that is about to be posted can be found in the system, the respective metadata is suggested to the active user. A possible extension of this feature, relying on social ties, would be to highlight metadata that has been entered by friends or members of the active user's groups. Another feature that relies on user networks is a discussion feature where users can discuss online with their peers. Often, users will rather not make their opinion about their colleagues' work completely public. However, when they review publications, they might be willing to share them with persons from their own network. Moreover, users can send interesting publications to other users. This enables another form of retrieval in tagging systems that is driven by actively sharing (in contrast to just posting a resource, which can be seen as a more passive form of sharing).

(RQ2) *The Personal Management Aspect:* We observed that users did not retrieve their own resources and tags as much as one would expect. A consequence, we already mentioned above, could be reminders of unvisited resources or clean-up recommendations to remove unused resources or tags and, thus, to keep the own collection manageable. Moreover, the observation suggests that visits to resources or tags could be valuable indicators for their importance to the user or generally to the system. In altmetrics – the study of measuring publications' impact through usage on the web – it is often assumed, that visits, downloads, and so on, are indicators for a publication's importance, and we will discuss correlations with citations in the next chapter. A consequence for webmasters of tagging systems would be to make use of these statistics (i) by showing them to the users in the system and (ii) by exploiting them in ranking and recommendation algorithms. Particularly the latter opens the field for new studies of recommender systems as almost any known folksonomic recommendation algorithm could be revisited and extended to make use of usage data.

(RQ3) *The Equality Aspect:* We found a strong *inequality* between the use of users, tags, and resources for navigation within BibSonomy. User pages are visited much

more often than resource or tag pages, providing clear evidence that assuming tags, users, and resources to be equally important for the navigation in BibSonomy would be wrong. This observation gives rise to a series of further investigations regarding the choice of the visited users during navigation. One could learn relations between the users and this information could be used to recommend users to visit, or even to recommend directly these users' resources, such that the active users no longer have to visit user pages to retrieve resources.

We also noticed a pronounced difference between retrieval of own content and content of others: Users tend to use the full-text search for the latter task, while they use tags to retrieve their own resources. This suggests that users know their way within their own vocabulary. On the other hand, since the full-text search in BibSonomy includes the full metadata of a publication as well as the tags, it makes sense for users to use it for finding new content, allowing hits to match the queries terms in the metadata as well as in tags. An idea for making tag search more successful for finding other users' content might be to extend the tag search beyond the Boolean approach of returning only posts that have the queried tag: Given a tag, one could return posts with similar tags. Hereby, similarity can be obtained from known word ontologies, like WordNet,<sup>29</sup> from semantics that are extracted from the tagging data in the system, from the "feedback cycle" of tagging – users find a post and then use their own tags when they copy it [Halpin et al., 2007] –, or from relations found in the log data (users who queried for tag X clicked on posts with tag Y, etc.). To gather further evidence for the behavior of users with tags of others, it would be helpful to repeat this study on a tagging system where the resources are not suitably describable by full-text – and, thus, a full-text search would be less helpful –, for instance, on a tagging system for images, like Flickr.<sup>30</sup>

Transition probabilities showed, that users often tend to stay with the same type of retrieval (e.g., one tag request after another). Requests to users are those with the highest share of different follow-up requests, suggesting that navigation by user leads to interesting resources or tags. Like the usage statistics, also the transition probabilities could be used as additional information in recommendation and ranking algorithms.

(RQ4) *The Popularity Aspect:* Finally, we compared popularity of entities in posts and in requests. We observed common usage patterns on an aggregate level, yet, the patterns are less pronounced on an individual level, suggesting that an entity's popularity in posts is only reflected to a certain extent in the requests to that entity. Such information could be valuable in the visualization of folksonomy information. Until now, tag clouds are the most popular means of displaying tags and usually, tags are ordered by their number of occurrences in posts. Our results suggest, that including the occurrences in requests could be helpful.

---

<sup>29</sup><https://wordnet.princeton.edu/>

<sup>30</sup><https://www.flickr.com/>

Overall, this chapter contributes a stepping stone for studies of social tagging systems by using actual traces of user behavior that can be found in request log data, a basis for comparative studies, exploring the extent to which these different aspects are pronounced in different tagging systems, and new insights about the use of literature in a publication management system.

### **5.6.1 Future Research**

It is reasonable to assume that different tagging systems (such as Flickr, Delicious, BibSonomy and others) exhibit unique characteristics and dynamics that make them amenable to different uses and purposes (see Section 5.2.2). Further studies of request log data in other tagging systems would be helpful in uncovering these differences. Finding that the equality assumption does not hold generally has important implications for the layout of tagging systems and for the design and implementation of algorithms that address search and retrieval. New approaches might incorporate actual transition probabilities and also consider the differences in popularity in posts and requests.

We hope our work triggers a new line of research on social tagging systems that utilizes traces of actual user behavior, to test and challenge our existing body of knowledge about these systems gained from other inquisition methods, such as surveys or post data.

Finally, request logs present a valuable resource for altmetrics studies, that is, studies on the usage intensity of resources. We will discuss the potential of BibSonomy for altmetrics and possibilities for predicting publication impact from data observed in a publication management system in the next chapter.

## Chapter 6

### Analyzing Publication Usage and Citations



Scholarly success is traditionally measured in terms of citations to publications or derivative metrics thereof. With the advent of publication management and digital libraries on the web, data from the usage phase of the scholarly publication life cycle has become a target of investigation, and new impact metrics computed on such usage data have been proposed – so-called altmetrics. Social bookmarking systems for scholarly publications allow their users to collect and manage publication metadata. By using such a system, researchers reveal their interest in the publications they post, visit, or export. In this chapter, we compare citations with altmetrics in the popular social bookmarking system BibSonomy. Our analysis, using a corpus of more than 250,000 publications, reveals that overall, citations and altmetrics in BibSonomy are mildly correlated. Furthermore, grouping publications by user-generated tags results in topic-homogeneous subsets, that exhibit higher correlations with citations than the full corpus. We find that posts, exports, and visits of publications are correlated with citations and even bear predictive power over future impact. A Random Forest predictor outperforms the baseline on average by about seven percentage points.

The analysis performed in this chapter is based on previously published material [Zoller et al., 2015, 2016].

### 6.1 Introduction

In this chapter, we continue our investigation of the usage of core features in a tagging system. After analyzing four different usage aspects in the previous chapter, here, we approach a fifth – the “altmetrics aspect”: The usage intensity of a resource can be interpreted as a measure for its relevance. Such information can be used in a tagging system to highlight particularly popular resources. Especially for the case where the resources are scholarly publications, the usage intensity could also be understood as a proxy for a publication’s impact.

Traditionally, the impact of a scholarly publication is estimated from its citations. However, that has the drawback that the results are only available long after an article has been published – simply because it takes time to write and publish new articles with a corresponding reference. Thus, citations can be used as impact indicators, but

they do not help researchers find papers which will be important for their discipline in the near future (e.g., within one year). With the advent of the social web, most scholarly communication and parts of the publication process have moved to the web and have thus become observable, among others in scholarly social bookmarking systems. Similarly to citing a publication, also storing it in an online reference manager or mentioning it in discussions or tweets can be regarded as an indicator for the publication's impact. This form of feedback is available more immediately – a new publication can already be bookmarked or tweeted about while it is being presented at a conference.

The creation of impact measures using such indicators on the social web – bookmarks, tweets, blog posts, and so on – has been subsumed under the umbrella term *altmetrics* (alternative metrics). It describes “the creation and study of new metrics based on the Social Web for analyzing, and informing scholarship”.<sup>1</sup> The Altmetrics Manifesto [Priem et al., 2010] explains the goals of this initiative, among them diversity in measuring impact, supplementing peer-review, and speed of availability. While altmetrics are meant to complement traditional citation counts, it is still relevant to study to which degree they are correlated with citations. Thus, in the last part of the manifesto, its authors note: “Work should correlate between altmetrics and existing measures, predict citations from altmetrics and compare altmetrics with expert evaluation.” This appeal was repeated recently by Bornmann [2014a] who listed missing evidence as one of the (current) disadvantages of altmetrics. Following this demand, in this chapter, we focus on the investigation of correlations between usage metrics and citations in BibSonomy (see Section 2.3.2), as well as on the predictive potential of usage metrics over citations.

Besides contributing to the altmetrics discourse and adding BibSonomy to the pool of web systems that can be used for altmetrics, our goal is to identify metrics which can support users of BibSonomy in finding relevant and high-impact literature, for instance, by implementing appropriate ranking and recommendation approaches. We determine correlations between citations (gathered from the scholarly search engine Microsoft Academic Search<sup>2</sup>) and several metrics that can be computed from the corpus of user-generated content (the bookmarked publication references) and the traces of usage behavior that are stored in the web logs of such a system. Previous studies usually focused on posts (i.e., the number of times that a publications has been added by the system's users). However, since we saw in Chapter 5 that popularity in posts and in requests is not strongly correlated, and particularly not for the less often used entities, next to posts, we investigate metrics based on requests, namely exports, visits, and requests to a publication's tags. All data is available within the system such that neither external sources nor the full texts of the publications are required for the computing the metrics.

---

<sup>1</sup><http://altmetrics.org/about/>

<sup>2</sup><http://academic.research.microsoft.com/>



After we compare citations to features on the full set of publications, we select subsets using the central feature in a bookmarking tool: tags. Using tags, we can group publications to topics without using external information about them, unlike previous works, where only publications of a particular conference or journal were chosen [Haustein and Siebenlist, 2011, Li et al., 2012, Saeed et al., 2008, Priem et al., 2012, Bar-Ilan et al., 2012] or where external classification systems were used to partition articles into disciplines [Haustein et al., 2014a, Thelwall and Sud, 2015]. Furthermore, we move beyond the analysis of correlations and approach the actual prediction of future citations. While we do not expect that data from BibSonomy alone is able to accurately predict citation counts (after all, BibSonomy is only one among many means to manage publications), we think it is important to analyze whether the observable usage data bears some predictive power over future citations and thus helps understand the form of impact that is measured by usage metrics.

**Research Questions.** Our research questions are the following:

- (RQ1) Despite our large corpus spanning various disciplines and publications of different quality, can we still detect a usage bias towards highly cited publications (in terms of correlations) in BibSonomy?
- (RQ2) Can the bookmarking system’s most inherent feature, tagging, be used to create topic-homogeneous subsets in which altmetrics exhibit higher correlations with citations than the full corpus?
- (RQ3) Do the observable traces of user behavior in the bookmarking system bear the potential to predict future citations (citations that occur *after* the observed usage of a publication)?

**Contributions.** In this chapter, we compare citations of publications to data from the usage phase of publications, gathered from BibSonomy. Answering the three research questions above, we go beyond previous work in the area of altmetrics

1. by explicitly comparing behavioral features to citations that occur in the near future (within one year) instead of comparing to all citations,
2. by using algorithms from machine learning to estimate the predictive potential over future citations,
3. by comparing more than one usage statistics (posts, views, exports, queries) in the above two tasks on a large dataset of more than 250,000 publications of multiple disciplines, and
4. by adding the use case of BibSonomy to the set of social web systems that have been investigated as possible sources for altmetrics.<sup>3</sup>

---

<sup>3</sup>We are aware of only one other altmetric investigation that included BibSonomy data: Haustein and Siebenlist [2011] computed metrics on the journal level, whereas we focus on the article level.

**Limitations.** The main limitation of this chapter is the restriction to usage metrics from BibSonomy. This is, however, unavoidable, as similar log data is hardly available from other systems. Further limitations arise from the choice of the citation dataset and from the experimental setup as a data-driven analysis. We discuss these in detail in Section 6.2.3.

**Structure.** In the next section, we explain the extracted features (the altmetrics in BibSonomy) and we discuss expectations and limitations. We then turn to related work in Section 6.3. We describe the dataset in Section 6.4 and present the analysis and findings regarding the above research questions in Section 6.5. Section 6.6 concludes the chapter.

The results of this chapter have previously been published in [Zoller et al., 2015] and are part of an extended version [Zoller et al., 2016]. All results presented here can also be found in the latter publication, but have been rearranged for this thesis.

## 6.2 Alternative Metrics in BibSonomy

In this section, we explain the article-level metrics we gather from BibSonomy. Furthermore, we explain how we count *future citations*. We then explain the tasks of determining correlations and predicting citations in Section 6.2.2, and we discuss expectations and limitations of our study in Section 6.2.3. For more information on BibSonomy and its features, see Section 2.3.2.

### 6.2.1 Usage Metrics and Future Citations

In our experiments, we use six different metrics as indicators for a publication's impact:

1. The metric  $post(p)$  counts how often a publication  $p$  was bookmarked. This is the same metric that was used in previous literature [Haustein and Siebenlist, 2011, Li et al., 2012, Saeed et al., 2008, Priem et al., 2012, Bar-Ilan et al., 2012].
2. With  $view(p)$ , we denote how often a publication  $p$  has been viewed (e.g., the publication's details page or a page with all posts about this publication from different users).
3. We denote with  $exp(p)$  the number of times a publication  $p$  has been exported into citation formats (e.g., BIB<sub>T</sub>E<sub>X</sub> or EndNote).
4. Since BIB<sub>T</sub>E<sub>X</sub> is the most often requested export format on BibSonomy, we additionally use the metric  $exp_{Bib}(p)$  to count exports of a publication  $p$  to that format.
5. We use  $req(p)$  to count all requests to a publication  $p$ , exports or otherwise, thus including the counts of  $view(p)$  and  $exp(p)$  in this metric.
6. Publications must be tagged in BibSonomy. With  $tag(p)$ , we count for a publication  $p$  how often one of its tags has been used in a search query.

Each metric is computed per publication and year. Hence, we can examine usage and citations both in individual years and over the total time (simply by adding up the respective metrics). The splitting by year also gives us the opportunity to compare the usage in the bookmarking system to citations in the future. In contrast to *early* citations, which refer to citations received shortly *after the publication* of a paper, with *future* citations, we refer to citations a paper receives *after some observed activity* related to that paper in the social bookmarking system. We must fix a time-frame in which we count such future citations. We decided to use the span of one year for two reasons: (i) Brody et al. [2006] compared download statistics of preprints on arXiv to future citations and found that after six months, this correlation was already high and increased only little if the delay was increased to one or even two years. Thus, six months would be a plausible option. However, since for the citing papers, the only available information about time is the year they were published – that is, the year in which they cited the publication at hand –, one year is the shortest time frame possible. (ii) The span of one year reflects the idea of the “hotness” of a paper and the ability to predict which publications will be highly cited within the following year would be valuable to researchers planning their next submissions. When we distinguish between citations in different years, we use the following convention: In general, citations are denoted with *cit*. Given an activity (e.g., *view*) to a publication in one year, we denote the number of citations to the publication within the same year by  $cit^{+0}$  and the number of citations within the next year by  $cit^{+1}$ . Thus  $cit^{+0}$  and  $cit^{+1}$  count disjoint subsets of the overall set of a publication’s citations ( $cit^{+0}, cit^{+1} \leq cit$ ).

## 6.2.2 Correlations and Prediction

In this chapter, we conduct two kinds of analyses: We measure correlation between usage metrics and citations, and we investigate the predictive potential of these metrics over citations in the future. For correlations, we report Pearson’s correlation coefficient  $r$ , as well as Spearman’s ranking correlation  $\rho$  (see Section 2.1.1). The latter has the advantage that it is suitable for non-linear relationships. Analogously to Haustein and Siebenlist [2011], we report both correlation coefficients, and we focus on  $\rho$  for the discussion.

To measure predictive power, we employ machine learning algorithms for classification. Classifiers are algorithms that automatically label given entities based on these entities’ features. The classifier computes a label (class) choosing from a previously fixed set of labels (classes). A classifier must learn how to pick a label for a given entity. In a training phase, the classifier is given a labeled dataset (i.e., entities with their features and their classes). The trained model can then be evaluated by applying it to an unlabeled dataset.

In our setting, the entities are publication-year pairs, and the features are the observed usage metrics in that year. Given a set of publications together with their

usage metrics<sup>4</sup> per year, we use all publication-year pairs  $(p, y)$ , where for publication  $p$  at least one of the three metrics was positive in year  $y$  (i.e., where the publication was used at least once in that year). We divide these pairs into two classes based on the number of citations in year  $y + 1$  using a threshold  $\tau$ : One class contains all publication-year pairs  $(p, y)$  where  $cit^{+1}(p, y) < \tau$  and the other class those where  $cit^{+1}(p, y) \geq \tau$ . For the threshold  $\tau$  we select the median of the number of citations per year (to publications in the subset at hand). Where the median was 0, we used  $\tau = 1$ . Thus the prediction task can be roughly summarized as: Given the usage of publication  $p$  in year  $y$ , predict whether  $p$  will have a higher impact, in terms of citations in year  $y + 1$ , than half<sup>5</sup> of the publications in the set.

In our experiments, we split the data into two sets: The *training set* for the classifiers contains the publication-year pairs of the years 2006 through 2008 (and thus the citations of the years 2007 through 2009). The *test set* contains the remaining pairs with usage features from 2009 and their citations from 2010. To evaluate the predictive power for a given classifier, the predicted classes (the results of the algorithm) are compared to the actual classes. We evaluate the result by its classification *accuracy* (*acc*), which is the share of correctly predicted entities.

**Example.** When an article  $a$  has been published in 2000, was posted in BibSonomy  $p_{2007}$  times in 2007, has been viewed  $v_{2007}$  times in 2007 and  $v_{2008}$  times in 2008, and has been exported  $e_{2008}$  times in 2008, it would yield the following publication-year pairs:  $(a, 2007)$  and  $(a, 2008)$ . Let us assume further that  $a$  has been cited  $c_y$  times in the year  $y$  (where  $y$  might be any year). Then our training dataset would contain the following data:

$$(a, 2007) : \text{post}(a) = p_{2007}, \text{view}(a) = v_{2007}, \text{cit}^{+1}(a) = c_{2008}$$

$$(a, 2008) : \text{view}(a) = v_{2008}, \text{exp}(a) = e_{2008}, \text{cit}^{+1}(a) = c_{2009}.$$

If  $a$  has been used in 2009 as well, say exported  $e_{2009}$  times, and cited by  $c_{2010}$  publications, then the test set will contain the data:

$$(a, 2009) : \text{exp}(a) = e_{2009}, \text{cit}^{+1}(a) = c_{2010}.$$

The classifier would try to predict whether  $c_{2010} > \tau$ , that is, whether the number of citations to  $a$  in 2010 will be larger than the average number of citations in 2010 for publications in the same subset as  $a$ . The prediction is then compared to the actual class obtained by the value of  $cit^{+1}$ . Note that publications in the test set can (but do not have to) occur in publication-year pairs of the test set and in pairs of the training set or in pairs of just one of these sets. The test set contains all pairs  $(a, 2009)$  for articles  $a$  that have been used at least once in 2009. The training set contains pairs  $(a, y)$  with  $y < 2009$ . Thus it is ensured that both sets are disjoint. Also note that the number of citations in the current year is not a feature used in the prediction. Only usage observed in BibSonomy is used as input for the classifiers.

---

<sup>4</sup>In Section 6.5.3, we will use the metrics *post*, *exp*, and *view*, following the results in Section 6.5.2.

<sup>5</sup>Since many publications receive equally many citations in a year, the classes are not exactly equally sized, depending on how many publications share the median.

As classifiers we selected implementations of *Random Forest* [Breiman, 2001] and *SVMs* with different kernels [Cortes and Vapnik, 1995], covering the two best classifier families at the moment (e.g., see [Fernández-Delgado et al., 2014]). For Random Forest we used the implementation of the R-package `randomForest`<sup>6</sup> with its standard configuration and with 100 repetitions per experiment. The SVMs were chosen from the *SVM*<sup>light</sup> package<sup>7</sup> using a radial and a polynomial kernel, again with default parameters.

### 6.2.3 Expectations and Limitations

Before we report the results of our analysis in Section 6.5, we discuss our expectations and also limitations of this study.

#### Expectations

In Section 6.2.1 we introduced five new metrics that complement the counting of posts (*post*) which has been investigated in previous studies. It is unclear whether these new measures will exhibit similar correlations with citations. The metrics *exp* and *exp<sub>Bib</sub>* cover the exports of publications to citation formats. Therefore, it is plausible that at least these two metrics would exhibit correlations with citation counts. In Section 6.1, we have explained that the dataset in this study is less restricted than those of previous studies – we review studies with different restrictions in Section 6.3. It contains arbitrary publications contributed by BibSonomy’s users instead of only publications from a particularly popular venue.

Furthermore, compared to similar publication management systems, BibSonomy belongs to the smaller systems: The largest currently available such system is Mendeley, which claims to have about four million users<sup>8</sup> and almost 100 million documents.<sup>9</sup> Probably the most similar to BibSonomy is CiteULike, which has more than eight million articles<sup>10</sup> at the time of writing. The current publicly available dataset<sup>11</sup> contains 145,744 users. BibSonomy currently has about 4 million (different) publications and about 3 million users of which 21,600 are classified as non-spammers with at least one publicly visible post. As a consequence, many publications are bookmarked by only one user and thus, the above described metrics yield low scores for many publications. Moreover, the publications in our corpus are distributed over various disciplines and hence over different publication and citation cultures. The dataset contains articles of various venues and different publication types (articles, conference or workshop contributions, preprints, and so on). It is well known in bibliometrics that

---

<sup>6</sup><http://cran.r-project.org/web/packages/randomForest/index.html>

<sup>7</sup><http://svmlight.joachims.org>

<sup>8</sup><http://blog.mendeley.com/elsevier/mendeley-and-elsevier-2-years-on/> (accessed August 28, 2015)

<sup>9</sup><https://www.mendeley.com/compare-mendeley/> (accessed August 28, 2015)

<sup>10</sup><http://www.citeulike.org/> (accessed August 28, 2015)

<sup>11</sup><http://www.citeulike.org/faq/data.adp> (accessed August 28, 2015)

both the scientific discipline and the venue are influential factors for a publication's probability of receiving citations (Bornmann and Daniel [2008] survey a variety of studies regarding these influences). We therefore expect much lower correlations than those reported in previous experiments (we will mention several in our review of related work in Section 6.3), and it is an open question whether there are relevant observable biases at all in BibSonomy.

By analyzing future citations – comparing usage in one year to citations in the next year – instead of just citations in general, we introduce another new aspect. It is unknown how that will influence the observable correlations. The hypothesis that users use tools like BibSonomy to manage those publications they plan to cite, is plausible as this is a main purpose of such a system. However, the reasons why users choose to post a publication are diverse: In BibSonomy we noticed that many users store metadata of their own work, possibly for representative or reporting purposes. Posting work of other authors might be for citing it later, but could also be just a reminder for “literature to-read”. Even papers that were meant to be cited when they were posted must not necessarily be actually cited in the final publication. For our analysis this means that we cannot expect to see posting a publication (and similarly viewing or exporting it) as direct indication of a new citation. Moreover, the publication management system which we investigate is only one among many tools to organize literature and to prepare an article's references section. Researchers may choose a different bookmarking system, offline tools, or simply files with reference lists on their desktop. Therefore, our system's user data covers only a small part of the worldwide process of scientific writing and thus of the creation of citations which are indexed by the search engine Microsoft Academic Search.

Furthermore, we will use tags to distinguish various topics and then investigate correlations on subsets of publications belonging to these tags. We expect that correlations will benefit from such restrictions since it narrows down the disciplines covered by the set of publications.

### **Limitations**

Our study is limited to the scholarly bookmarking system BibSonomy. Similar data on the usage of a comparable system is simply not available, especially the web server logs, which contain sensitive information about the system and its users. We can speculate about results on other systems: Priem et al. [2012] and Li et al. [2012] compared the bookmarking systems CiteULike and Mendeley (each using a different set of publications) regarding correlation between the number of posts and the number of citations. Both found consistently that correlations were higher on the larger system, Mendeley. Li et al. [2012] also observed high correlations between the post counts in Mendeley and CiteULike. We therefore hypothesize that, similarly, smaller (larger) systems than BibSonomy would exhibit similar or lower (higher) correlations.

To count the citations, we used Microsoft Academic Search and we discuss limitations of this data source in the next section. Again, the bottleneck is the availability of

data from other sources (especially in the large quantities required in this study). Orduna-Malea et al. [2014] and Haley [2014] observed high correlations between Microsoft Academic Search and Google Scholar for various metrics, Li et al. [2012] observed high correlations between Google Scholar and the Web of Science. We thus can assume that the results in our experiments would be similar if we had used another valid source for the citation counts.

The analyses presented here are driven by the data of the social web system in which we gather the altmetrics. Thus, only publications that occur at least once in BibSonomy are included – obviously a small subset of the complete body of scholarly publications. In the notions of Costas et al. [2014], this constitutes a *tight* analysis, while others called it a *non-zero analysis* (e.g., [Mohammadi and Thelwall, 2014]). Similarly to [Waltman and Costas, 2014], it assumes the point of view of the bookmarking system’s operators who can only observe the activities in their system. An alternative approach would have been to use some other body of literature that contains “cited-by” information and set the respective usage metric of BibSonomy to zero if the publication is not covered in BibSonomy. Due to the relatively small size (e.g., compared to the publication management system Mendeley, see our comparison above) such a dataset would probably be dominated by zeros on the BibSonomy side. However, such a corpus was not available to us and for the goal of predicting citations for publications in BibSonomy, the chosen approach is preferable.

There are also some limitations of altmetrics in general, pointed out, for example, in [Wouters and Costas, 2012], that apply for BibSonomy and for our study as well: As data is user-generated, it is error-prone and it depends on the kind of users the system attracts. Research disciplines have different practices regarding citing or discussion literature and thus multidisciplinary studies are difficult. We address this particular challenge in Section 6.5.2, where we use tags to produce topic-focused subsets of publications. We also review risks and opportunities of using altmetrics and usage-based indicators in general to assess a publication’s quality in Chapter 9. Finally, BibSonomy is only one system and thus the coverage of available publications (which can only be guessed) is low. Still, as we will show, biases towards more often cited publications exist.

## 6.3 Related Work

In this section, we review some literature on scientometrics and altmetrics on the web and particularly previous work that has dealt with the use of social bookmarking systems to assess scholarly impact. We compare our work to previous experiments and outline the differences between those approaches and ours.

The problem of availability of citation-based impact measures has been mitigated through web search engines like Microsoft Academic Search or Google Scholar which compute such metrics on publication data crawled from the web. Fu et al. [2014] demonstrated with their system *pubstat.org* how such data can be used to compute

various rankings of publications, authors, or venues. Such measures can even be used within other tools that scientists use for their research. A practical example is *Scholarometer* [Kaur et al., 2014] which allows users to describe (tag) authors and compute impact statistics using data from Google Scholar. However, these tools suffer from the drawback of citations being available only a rather long time after an article’s publication. The idea of altmetrics is to create impact measures that are available much faster, by deducing impact from the usage of publications in web-based tools themselves. For example, Mas-Bleda et al. [2014] showed for a set of highly cited researchers that among those who made active use of the social web for sharing publications or slides, almost all created some form of impact, measured in terms of document or profile views in these systems. Next to the aspect of speed, other altmetrics have other advantages, like diversity, openness, and broadness [Bornmann, 2014a]. Broader applicability of altmetrics, beyond measuring scientific success, was, for example, demonstrated by Bornmann [2014b], who investigated connections between both altmetrics and citations to societal impact, indicated by experts of the publishing and peer-reviewing platform F1000 (Faculty of 1000). Among others, it was found that publications with the tag “good for teaching”, which indicates a certain relevance for non-researchers, created more impact in altmetrics and particularly in Twitter counts than publications without that tag. Another advantage of altmetrics could be that they allow measuring the impact of other scholarly output that is rarely cited, like research datasets. However, Peters et al. [2015] found, by comparing citations and coverage in three altmetrics aggregators for datasets covered in the Web of Science that this is not (yet) the case.

Good starting points for literature on altmetrics in general are its manifesto [Priem et al., 2010] as well as the altmetrics workshops.<sup>12</sup> Like suggested in the manifesto, several experiments have shown correlations between the usage of a publication in a web system and the number of citations to that publication. For example, Brody et al. [2006] showed that download counts of articles (taken from the section of high energy physics on the preprint server arXiv) correlate well with later citations if downloads are counted over a period of at least six months (reported is Pearson’s  $r = 0.397$ ). Eysenbach [2011] compared citations and tweets to 55 articles of the Journal on Medical Internet Research in a number of experiments. He found correlations between citations to an article (recorded on Google Scholar and Scopus) and the number of tweets to an article. Counting only tweets from within seven days after an article’s publication, he observed correlations of  $\rho = 0.36$  (with citations in Google Scholar) and  $\rho = 0.22$  (Scopus). Among others, he also noticed that the 25 % articles with the highest number of tweets within seven days contained 75 % of the 25 % articles with the highest number of (subsequent) citations. He noted that the articles with many tweets but few citations were mostly relevant for patients (rather than for researchers) and thus had a different form of impact that is usually not acknowledged

---

<sup>12</sup><http://altmetrics.org/workshop2011/>, <http://altmetrics.org/altmetrics12/>, <http://altmetrics.org/altmetrics14/>



in citations. Thelwall et al. [2013] used a set of PubMed articles to compare their citations to metrics counting the activities in eleven different social web systems – among them Facebook, Twitter, Reddit, and LinkedIn, as well as forums and blogs. They found evidence for an association between usage and citations for six of these systems, but also that for each system, the share of articles from their corpus that was used there, was rather low (below 20%). Similarly, Haustein et al. [2014a] compared the microblogging system Twitter to the publication management system Mendeley on a corpus of 1 million bio-medical articles from the Web of Science and from PubMed. They found the two altmetric sources to be very different in terms of coverage and correlation with citations, measured for 13 disciplines. They concluded that altmetrics from different sources reflect different forms of impact; for example, Mendeley might reflect academic impact, while altmetrics on Twitter could be an indication for impact in the broader public. The latter was supported by Haustein et al. [2014b].

Investigations on the author level have been conducted by Ortega [2015]: Medium to high correlations were found between (publicly available) usage measures from ResearchGate and citations from Microsoft Academic Search and from Google Scholar, while between citations and social measures<sup>13</sup> in Academia.edu, Mendeley, or ResearchGate, only a small or no correlation was observed.

In contrast to the works mentioned above, in this work, we focus on scholarly social bookmarking, since among the platforms on the web where researchers exchange thoughts, opinions, and ideas, these systems are the ones that are dedicated directly to managing and sharing publications. Platforms like Twitter or Facebook have a much broader scope. This intuition was, for example, confirmed by Priem et al. [2012], who found that for a given set of publications, dedicated scholarly bookmarking systems (in their study Mendeley and CiteULike) had a much higher coverage of these publications in bookmarks (about 80% on Mendeley and about 31% on CiteULike) than other web platforms like Facebook, Twitter, or Wikipedia (all less than 15%). Furthermore, we focus on the level of individual publications rather than on authors or venues.

### 6.3.1 Measuring Scholarly Impact in Social Bookmarking Systems

Tagging and managing scholarly content is the key functionality of publication bookmarking systems. These systems, like BibSonomy, CiteULike, or Mendeley, allow their users to create collections of publications and to annotate each publication with a set of tags. Although posting or visiting a publication does not automatically imply a later citation, it can still be regarded as an expression of interest in that publication. Correlations between publications' citations (recorded either in expert-controlled databases like the Web of Science or in corpora of crawled documents from the web like Microsoft Academic Search) and their occurrence in social bookmarking systems have been analyzed before on different levels: on the level of journals [Haustein and Siebenlist, 2011], on the level of authors [Bar-Ilan et al., 2012], and on the level of

---

<sup>13</sup>Social measures can be computed as altmetrics on the author level; Ortega [2015] considered followers and followees per author.

publications [Li et al., 2012, Saeed et al., 2008, Bar-Ilan et al., 2012, Priem et al., 2012].

Haustein and Siebenlist [2011] investigated correlations between the use of journal articles in social bookmarking systems and several journal-level citation indicators. They found medium to high correlations (Spearman's correlation, ranging from  $\rho = 0.240$  to  $\rho = 0.893$ ) between the number of bookmarks to articles of a journal and various journal-level metrics. They used a dataset spanning 45 solid state physics journals and bookmarks from three bookmarking systems (including BibSonomy). Bar-Ilan et al. [2012] investigated the social bookmarking systems Mendeley and CiteULike and compared the number of posts to the number of citations, recorded in the publication database Scopus. Using a total of 1,136 articles – a sample generated as the set of all publications of 57 authors who had attended the conference STI 2010 – they found medium correlations ( $\rho = 0.232$ ) between CiteULike and Scopus and higher correlations ( $\rho = 0.448$ ) between Mendeley and Scopus. Saeed et al. [2008] conducted an experiment on the 84 publications of the conference WWW 2006. They found a strong rank correlation ( $\rho = 0.6003$ ) between the number of citations and the number of bookmarks that a publication receives. This correlation is much stronger than that found for citations and a co-authorship-based ranking of the same publications. Similarly, Li et al. [2012] observed rank correlations between 0.304 and 0.603 between post counts in the bookmarking systems CiteULike and Mendeley and citations on the Web of Science for 793 Nature and 820 Science articles. Finally, Priem et al. [2012] examined citation counts from the Web of Science and bookmark counts on both Mendeley and CiteULike for articles from three PLoS ONE journals. They found ranking correlations of  $\rho = 0.3, 0.2, 0.2$  for CiteULike and  $\rho = 0.3, 0.5, 0.4$  for Mendeley, depending on the journal.

The above-mentioned studies [Li et al., 2012, Saeed et al., 2008, Bar-Ilan et al., 2012] considered only relatively small sets of publications. While the corpus of Bar-Ilan et al. [2012] comprised the oeuvres from the Web of Science and Scopus, the other four mentioned studies used sets of quality-homogeneous publications from high profile venues, such as particular conferences or journals. In contrast to these restrictions, our corpus includes any publication that users have posted to BibSonomy.

Thelwall and Fairclough [2015] used synthetic datasets to simulate publications with their citations and with ratings. They found strong differences between the correlations between citations and ratings on homogeneous datasets (each representing a single discipline) and heterogeneous datasets. The effect depends on the characteristics of the individual datasets that are merged (e.g., mean number of citations, correlations). Thelwall and Wilson [2014], Thelwall [2015] and Brzezinski [2015] confirmed that publication corpora of various disciplines indeed possess different statistical properties. The dataset in BibSonomy is per se heterogeneous as there is no mechanism in BibSonomy to explicitly assign a publication to some discipline. Thus, we can expect that our results on the full dataset will be lower than those reported in the studies mentioned above. Recently, a lot of research has been conducted on Mendeley, presumably due to its size (see below) and thus higher coverage of publications.

[Mohammadi and Thelwall, 2014, Mohammadi et al., 2015, Thelwall and Wilson, 2015, Zahedi et al., 2015] all compared Mendeley readerships to citations from either the Web of Science or Scopus using different sets of publications from different levels of aggregation by discipline. Mohammadi and Thelwall [2014] considered the social sciences and humanities and found Spearman correlations of 0.516 and 0.428 and slightly higher or lower values for their sub-areas. Mohammadi et al. [2015] used the main disciplines according to the US National Science Foundation classification and journal articles published in 2008. For all five disciplines, Spearman correlations between 0.501 and 0.561 were measured. Thelwall and Wilson [2015] computed correlations between Mendeley readerships and citations on 47 fields of medical research according to Scopus and found Spearman correlations between 0.379 and 0.784 for the individual fields and 0.697 for all medical publications together. Finally, Zahedi et al. [2015] followed the classification of the Leiden Ranking, which assigns one of five large research fields to any publication. On the overall dataset, a correlation of 0.52 was measured and values between 0.43 and 0.60 for the five fields.

For those parts of our experiments that use smaller subsets of the corpus, we exploit the posts' tags. Tags reflect the users' perspective on publications rather than that of a publisher or author [Peters et al., 2011]. Thus, we use the bookmarking system's intrinsic way of determining topic structures, rather than external knowledge about venues, and we do not restrict the corpus to publications of the same quality level. Furthermore, all studies above focused only on the visible representations of publication usage, namely their bookmarks (posts). In this chapter, for the first time, we will complement bookmark counts with other usage metrics that can be computed in a social bookmarking service.

Finally, all above-mentioned studies demonstrate medium to high correlations between the number of bookmarks or readerships and the number of citations to a publication. However, although the early availability of altmetrics is one of the key advantages of these measures, all these studies ignore time when they compute the correlations. In this chapter, however, we investigate particularly the correlation between usage metrics and citations that occur in the future.

Temporal aspects of altmetrics have been considered by Thelwall and Sud [2015], who compared citations counts from Scopus and readerships from Mendeley per publication year, for articles from 50 Scopus sub-categories. They found that correlations are stronger and relatively stable for publications that had been published five years ago or earlier and that citers accumulate slower than readers. A plausible conclusion they suggest, is that in the early years Mendeley readerships are valuable impact indicators. In our work, we go even further by measuring correlations directly between altmetrics in one year and citations in the future.

A variety of recent studies of the publication management system Mendeley have covered a number of further aspects beyond measuring correlations. For instance, Zahedi et al. [2015] investigated how correlations depend not only on the scientific field but also on the academic position of the users. In all five fields and in total, the usage metrics comprising only PhDs yield the strongest correlations, librarians

the lowest. They also studied the ability to filter the most highly cited publications using precision and recall on rankings in which they identified the most highly cited publications (according to the Web of Science). It turned out that rankings that order publications by their Mendeley readership, are better filters than rankings ordered by the journal citation score. The approach of prediction in this work is fundamentally different as it considers only citations occurring in the future (i.e., in the year after the measured use).

To discover limitations of the use of altmetrics as precursor for citations, Thelwall [2015] selected outliers in 15 disciplines – publications that either had few citations but a large number of Mendeley readers, or vice versa. Using human judgment, various reasons, technical (e.g., erroneous indexing) and legitimate (e.g., publications are interesting to users who do not actively publish, reading does not imply citation, recipients did not use Mendeley) were identified. These same limitations apply to BibSonomy (particularly because it has fewer users (see below) and must be kept in mind when interpreting the results.

## 6.4 Dataset

For our experiments, we combine metadata on publications from the two web systems BibSonomy and Microsoft Academic Search. In the following, we first describe the two datasets and then a few challenges merging them.

### 6.4.1 BibSonomy

The dataset used in this chapter is created from both BibSonomy’s web server logs and database contents, spanning the time from 2006 (launch of the system) until the end of 2009.<sup>14</sup> In the data, each publication is identified through a hash value that is computed using its title, authors (or editors, when no authors are given) and publication year (see [Voss et al., 2009] for more details). To ensure that only requests from real users are captured in the usage data from the web server logs, we employed a heuristic filtering based on the HTTP request’s status code and referer header. We removed redirects that were automatically initiated by BibSonomy and not by the choice of the user (e.g., redirects to the user’s personal page after editing a post). Another heuristic was used to remove requests of bots (in particular crawlers from search engines), based on the request’s user agent header. We utilized well-known user agents’ strings from various online sources, as well as user agents of clients which showed abnormal request behavior. The remaining dataset contains about 40 million requests in the considered period. Anonymized datasets of logs and posts are made available to researchers by the BibSonomy team.<sup>15</sup>

---

<sup>14</sup>Both datasets also include data from later years, but for our experiments, we had to restrict them (see Section 6.4.3).

<sup>15</sup><http://www.kde.cs.uni-kassel.de/bibsonomy/dumps/>

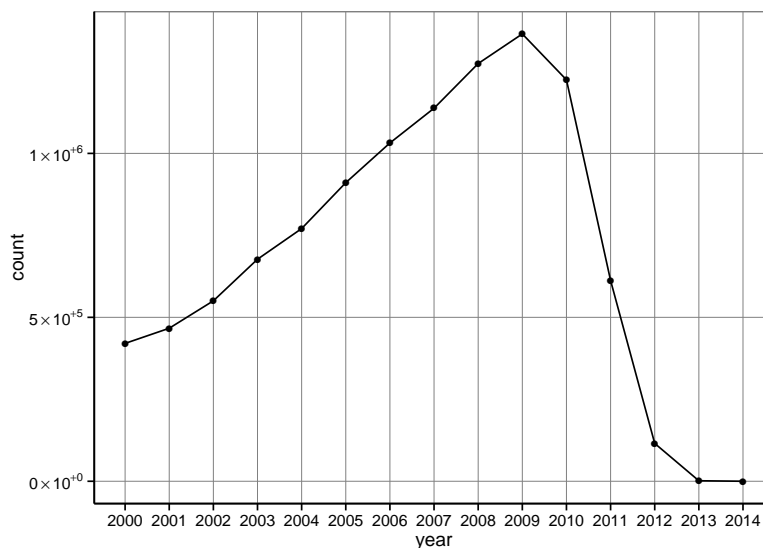


Figure 6.1: Citations to publications in BibSonomy according to MAS, distributed over the years. For better readability only the last 15 years are displayed.

### 6.4.2 Microsoft Academic Search

Microsoft Academic Search (MAS) is a web search engine that indexes research literature and their citing publications (i.e., publications that reference another paper). Similar to the service Google Scholar, publication data is obtained by crawling the web and extracting information from publications. According to Khabsa and Giles [2014], MAS contains roughly as many records as the Web of Science. While it was shown that Google Scholar is superior to MAS, especially in terms of covered publications, it was also found that for computer science (which accounts for the majority of publications in BibSonomy) MAS even has a slightly higher coverage than Google Scholar [Khabsa and Giles, 2014, Orduna-Malea et al., 2014]. Thus, and due to restrictions in Google Scholar’s robots directives, we chose MAS to retrieve the required citation data for all publications in BibSonomy. We will discuss the issue of data availability in MAS and our adaptation of the dataset in the next section. In our study, we use all citations that MAS lists for a publication. Particularly, we do not remove self-citations for three reasons: (i) In Section 2.2.1, we already discussed the controversial nature of removing self-citations and that popular metric providers like the Thomson Reuters Journal Citation Reports or Google Scholar include them in their computations. (ii) Removing them would require rigorous manual cleaning to correct possible errors in the author names. While we chose to do this in Chapter 3 with a much smaller corpus, a similar effort in this study forbids itself due to the size of the corpus. (iii) If self-citations were ignored, one would also have to address self-usage, that is, users of BibSonomy who

post their own publications. Posting own work is common and reasonable practice, however, identifying these instances would require matching BibSonomy user names to authors, which is not possible using the available data.

### 6.4.3 Matching between BibSonomy and Microsoft Academic Search

Dealing with publication metadata is often tedious: The data from BibSonomy is user-generated and thus prone to contain spelling errors and incomplete attributes, making it difficult to match publications. However, also the data about references in publications – as it is collected and extracted by MAS – can be erroneous and often contains missing values (e.g., missing publication years, typos, etc.). Furthermore, any search engine will only cover a subset of the number of all existing publications (see also [Khabsa and Giles, 2014]) and thus there are publications in BibSonomy that cannot be found in MAS.

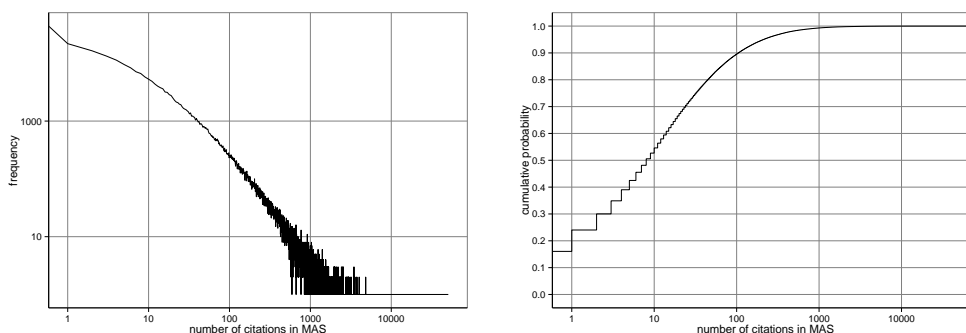
To collect information about the citations from papers in MAS to publications bookmarked in BibSonomy, we queried MAS for each such publication individually (by title and authors’ last names) and collected the top result together with all the citations that were registered for it. We excluded posts from bot users in BibSonomy (e.g., an importer mirroring the publication database DBLP). Since the top result did not always match the query, we applied the following pre-processing steps to ensure that the found publication corresponds to the queried publication: We compared the queried publication’s title with the found publication’s title by (i) removing whitespace, accents, and special characters like  $\text{\LaTeX}$  entities or punctuation and (ii) computing the Damerau–Levenshtein distance [Damerau, 1964], an extension of the well-known Levenshtein distance, that additionally allows transpositions of characters. We considered a publication to be a correct match, when the Damerau–Levenshtein distance was less than four. Thus, we neglect small typos (e.g., transpositions of letters), or missing articles “the” or “a”.

Of the 678,796 publications that had been posted to BibSonomy between 2006 and 2012 by regular users, for 279,321 we could find a corresponding publication in MAS (according to the rule above) when we crawled the service in early 2014. The reasons that many publications did not yield a result are many-fold and we here list those that became apparent by manually checking publications:

- Not all scientific publications are indexed by MAS.<sup>16</sup> Especially publications that appeared after 2010 are rarely indexed (see below).
- Not all publications in BibSonomy are scientific. Some users also bookmark belletristic literature, (non-academic) non-fiction like programming guides, blog posts, presentations, and so on.
- Not all publications in BibSonomy actually have been published – among them preprints, manuscripts, or bachelor/master theses.

---

<sup>16</sup><http://academic.research.microsoft.com/About/help.htm>



(a) The frequency distribution (visualized on a log-log scale). (b) The cumulative distribution function (visualized on a log-linear scale).

Figure 6.2: The frequency distribution and the cumulative distribution function of the number of citations recorded in MAS for publications in BibSonomy.

- The data in BibSonomy is user-generated. The publications are either entered manually, imported from other citation managers, or scraped from publisher websites. All three methods have their pitfalls and thus many publication titles have missing words or are abbreviated or spelled incorrectly.

Orduna-Malea et al. [2014] pointed out that the number of new publications indexed by MAS experienced a slight drop in 2010, a large drop in 2011, and an even larger drop afterwards. Therefore, we counted the number of citations to publications in our corpus. Figure 6.1 shows the number of publications in MAS per year that cited at least one publication in the BibSonomy dataset (for better readability only the last 15 years are plotted). The observed trend is very similar to what is described by Orduna-Malea et al. [2014] (especially to their Table 7). This means that (i) our subset of publications exhibits the same distribution of indexed citations as the full set of publications indexed in MAS, which was used by Orduna-Malea et al. [2014], and (ii) the sets of indexed publications from the years 2011 and later cannot be considered to be representative sources for citations. Because several experiments in this work focus on citations occurring in a particular time frame after an activity related to a publication was observed in BibSonomy, we decided to include only the years until 2010. Thus, when we compare features in BibSonomy in one year to citations in the following year, we can use BibSonomy features from the years 2006 through 2009 and citations from 2007 through 2010. After removing publications that appeared after 2009, 253,749 publications remain in our corpus.

#### 6.4.4 Citation Frequency Distribution

The frequency distribution of the total number of citations in MAS to publications in the BibSonomy dataset is shown in Figure 6.2(a). Similar frequency distributions

on other citation datasets were observed in previous works, such as [Redner, 1998]. Most of the publications in BibSonomy were not cited by any other scientific work and publications with only one citation represent the second largest subset in the crawled dataset. The frequency decreases continuously with higher numbers of citations, but also starts to oscillate for citation counts larger than about 100. Additionally, the cumulative distribution function is displayed in Figure 6.2(b). We can observe that more than half of the publications in the dataset were cited less than or exactly ten times. About 89 percent of all publications have less than 100 citations.

We fitted the citation distribution to a power law, a probability distribution that is proportional to a function  $x^{-\alpha}$  for all  $x$  above some threshold  $x_{\min}$ . To determine the optimal fit, we used the methodology by Clauset et al. [2009] (see Section 2.1.1), that was also used in previous studies of citation distributions [Brzezinski, 2015, Albarrán and Ruiz-Castillo, 2011]. The optimal fit has parameters  $\alpha = 2.59$  and  $x_{\min} = 808$ . The threshold  $x_{\min}$  is very high:  $x_{\min} = 808$  means that the part of the distribution that is fitted contains only publications with at least 808 citations. We therefore also consider the second (local) optimum:  $\alpha = 2.34$  for  $x_{\min} = 303$ . The exponents  $\alpha$  of both fits are lower than reported by both Brzezinski [2015] and Albarrán and Ruiz-Castillo [2011]. This is evidence that our corpus, which is collected from a web system, is indeed different to corpora that are collected from traditional article catalogs like Web of Science or Scopus. Compared to fits of other possible candidate distributions, we find that both power-law fits are better than fits to exponential distributions, but also that other heavy-tailed distributions (among them the lognormal distribution) provide better fits. In that respect, the distribution of citations in our corpus is consistent with Brzezinski [2015], who similarly observed that other functions fit the empirical citation distributions of various disciplines better than power laws, and with Thelwall and Wilson [2015], who observed for 45 medical sub-fields that lognormal distributions are better fits to citation counts than power laws. Thelwall and Wilson [2015] further showed that for these disciplines a hooked power law is an even better fit than the lognormal distribution.

Moreover, the high  $x_{\min}$  values suggest the presence of a “hooked power law” [Thelwall and Wilson, 2014], that is, a distribution proportional to  $(x + B)^{-\alpha}$ , where  $B$  is a parameter with  $B > -1$ . The parameter  $B$  causes the power law to shift along the  $x$ -axis. Thus with higher  $B$ , especially the values for small  $x$  are smaller than they would be in a plain power-law distribution. Regular power laws are hooked power laws with  $B = 0$ . Thelwall and Wilson [2014] showed that when the full distributions (including all citations from those publications that have been cited at least once) are fitted, a hooked power law with  $B > 0$  is a better fit than regular power laws. Indeed, fitting a hooked power law to the distribution of MAS citations for publications in BibSonomy, we find the parameters of the optimal fit to be  $\alpha = 1.93$ , and  $B = 10.16$ . These values suggest the presence of a hooked power law rather than that of a power-law distribution. It is noteworthy that the parameter  $\alpha$  is lower than those measured by Thelwall and Wilson [2014] for citation distributions of various research areas (the minimum  $\alpha$  there was  $\alpha = 2.51$ ).



Table 6.1: Correlation between different behavioral features in BibSonomy and the number of citations to a publication. The upper right triangle shows Pearson’s  $r$ , the lower left triangle shows Spearman’s  $\rho$ . All values are significant at the 0.01-level. Correlations are computed over all publications in the dataset.

	<i>post</i>	<i>view</i>	<i>exp</i>	<i>expBib</i>	<i>req</i>	<i>tag</i>	<i>cit</i>
<i>post</i>	1	0.644	0.638	0.633	0.446	0.330	0.181
<i>view</i>	0.322	1	0.725	0.705	0.656	0.322	0.091
<i>exp</i>	0.317	0.429	1	0.988	0.742	0.279	0.157
<i>expBib</i>	0.328	0.417	0.955	1	0.722	0.277	0.160
<i>req</i>	0.325	0.912	0.663	0.634	1	0.213	0.072
<i>tag</i>	0.277	0.272	0.237	0.242	0.267	1	0.036
<i>cit</i>	0.199	0.098	0.122	0.120	0.098	0.014	1

By and large, we conclude that the distribution in our corpus is qualitatively similar to citation distributions that have previously been analyzed, however quantitatively, there are pronounced differences (the exponent  $\alpha$  is lower than previously observed both in the power-law and the hooked power-law fit).

## 6.5 Analysis

In this section, we present the results of our study. We begin with experiments that analyze different features on the full corpus of publications over all years, before we focus on subsets in Section 6.5.2 and attempt the prediction of citations in Section 6.5.3.

### 6.5.1 Correlations on the Full Corpus

Our first experiment is similar to those in the literature mentioned in Section 6.3, as it ignores time; yet also different as it uses a much more inhomogeneous corpus mixing publications from different disciplines and quality levels. We compute correlations between behavioral features and citation counts over all publications in our corpus.

Table 6.1 shows Pearson’s  $r$  and Spearman’s  $\rho$  for each pair of metrics: We can observe significant positive correlations for each pair of behavioral features as well as between each such feature and the number of citations. As expected (see Section 6.2.3) the correlation between post counts and citations is lower than in previously reported experiments with strong restrictions on the set of publications. Yet, we still observe a small correlation that clearly indicates a bias in the behavior of users towards posting rather highly cited publications more often. Regarding the other behavioral features, we observe another noticeable bias between exporting (*exp*) and citing publications. The choice between all exports (*exp*) and BIB<sub>TEX</sub> exports (*expBib*) makes little difference –

Table 6.2: For each usage metric, the correlations (Pearson’s  $r$  and Spearman’s  $\rho$ ) between the metric and citations in the same year ( $cit^{+0}$ ) and citations in the following year only ( $cit^{+1}$ ). All values are significant at the 0.01-level. Correlations are computed over all publication-year pairs in which the publication has been used at least once (the according behavioral metric is non-zero). The number of such pairs is  $N$ .

metric	Pearson’s $r$		Spearman’s $\rho$		$N$
	$cit^{+0}$	$cit^{+1}$	$cit^{+0}$	$cit^{+1}$	
<i>post</i>	0.20	0.20	0.16	0.16	194,012
<i>view</i>	0.12	0.12	0.14	0.14	64,355
<i>exp</i>	0.17	0.17	0.15	0.15	33,989
<i>exp<sub>Bib</sub></i>	0.17	0.17	0.15	0.15	31,985
<i>req</i>	0.06	0.06	0.14	0.14	77,018
<i>tag</i>	0.04	0.03	0.07	0.06	399,502

both features are almost perfectly correlated. This can easily be attributed to the fact that `BIBTEX` is the most often used export format in BibSonomy. No real correlation can be observed between the *tag* metric and citation counts. A possible explanation for this lack of correlation is that one tag can occur in many posts and thus the metric is not publication-specific enough. Finally, apart from *exp* and *exp<sub>Bib</sub>*, and *req* and *view*, none of the behavioral metrics is strongly correlated with another one. Particularly between *post* and the other metrics we find medium correlations, indicating that while these metrics are not completely diverse, they are valuable complements to just counting posts.

In the next analysis, we additionally restrict the time in which a citation occurred and observe correlations between the behavioral metrics and either citations in the same year ( $cit^{+0}$ ) or citations in the next year only ( $cit^{+1}$ ). For that purpose, we use for each behavioral metric those publication-year pairs, where the metric is positive, that is, where the publication was used at least once in that year, according to the metric. Table 6.2 shows the results: All metrics except *tag* (like before) exhibit medium correlations with citations in the near future ( $0.14 \leq \rho \leq 0.16$ ). Correlations with citations in the same year and with citations in the next year are almost identical. This confirms the hypothesis of a bias in the usage behavior towards both publications that are already relevant and those that will be soon. The small correlations can be explained by the fact that our data contains results from multiple disciplines, which reduces correlation strengths [Thelwall and Fairclough, 2015]. Since no explicit classification by discipline is available for the publications in BibSonomy, we further investigate this issue in the next section by considering a grouping that is an integral part of social bookmarking, namely the tags.

Table 6.3: For each behavioral metric, the correlations (Pearson’s  $r$  and Spearman’s  $\rho$ ) between the metric and citations in the same year ( $cit^{+0}$ ) and citations in the following year ( $cit^{+1}$ ) – each averaged over the 30 tag-induced corpora. Correlations are computed over all publication-year pairs in which the publication has been used at least once (the according behavioral metric is non-zero). The average number of such pairs is  $N$ .

metric	Pearson’s $r$		Spearman’s $\rho$		$N \pm \text{sd.}$
	$cit^{+0} \pm \text{sd.}$	$cit^{+1} \pm \text{sd.}$	$cit^{+0} \pm \text{sd.}$	$cit^{+1} \pm \text{sd.}$	
<i>post</i>	0.32 $\pm$ 0.09	0.33 $\pm$ 0.10	0.29 $\pm$ 0.07	0.31 $\pm$ 0.07	1,551.0 $\pm$ 950.5
<i>view</i>	0.16 $\pm$ 0.08	0.17 $\pm$ 0.08	0.20 $\pm$ 0.06	0.24 $\pm$ 0.07	1,119.6 $\pm$ 516.8
<i>exp</i>	0.28 $\pm$ 0.14	0.28 $\pm$ 0.13	0.26 $\pm$ 0.07	0.27 $\pm$ 0.07	664.7 $\pm$ 316.5
<i>expBib</i>	0.28 $\pm$ 0.15	0.28 $\pm$ 0.14	0.26 $\pm$ 0.07	0.28 $\pm$ 0.07	634.7 $\pm$ 303.7
<i>req</i>	0.19 $\pm$ 0.10	0.20 $\pm$ 0.09	0.22 $\pm$ 0.06	0.25 $\pm$ 0.07	1,260.0 $\pm$ 579.6
<i>tag</i>	0.12 $\pm$ 0.09	0.12 $\pm$ 0.09	0.14 $\pm$ 0.09	0.13 $\pm$ 0.10	4,890.7 $\pm$ 3,651.5

### 6.5.2 Correlations for Popular Topics

It is well known in scientometrics (see Section 6.2.3) that different scientific communities have different publication and citation cultures, and that publications in more popular areas (hot topics or large research areas) often receive more citations than others. In a tagging system, one purpose of the tags is to indicate the topics of the bookmarked resources [Golder and Huberman, 2006]. It is thus natural to use these tags to group publications into topic subsets. For that purpose, we computed the 30 most popular tags, measuring a tag’s popularity as the number of users who used it at least once. We excluded stop-words and system tags (e.g., the tag “myown”), removed all characters that were neither numbers nor letters from the tag string, and used Porter’s stemming algorithm [Porter, 1980] to aggregate different occurrences of the same word stem (e.g., “algorithm” versus “algorithms”).

For each tag stem, we selected those publications that have been annotated with a tag having that stem at least once. We repeated the computations described in the previous section for each of the resulting 30 smaller corpora of publications, comparing the behavioral metrics to citations in the same year ( $cit^{+0}$ ) and one year in the future ( $cit^{+1}$ ). For the sake of legibility, we report averaged numbers together with their standard deviation in the following. However, overviews with all correlation values can be found in Appendix B (Table B.1 for  $cit^{+0}$  and Table B.2 for  $cit^{+1}$ ). Table 6.3 is similar to Table 6.2, only instead of correlations on the full corpus they are now measured on each of the 30 small ones individually and then averaged.<sup>17</sup> We note that these (unweighted) average correlations are all much higher than those observed on

<sup>17</sup>There are 720 individual correlation values (30 tag stems, 6 metrics, 2 citation measures ( $cit^{+0}$  and  $cit^{+1}$ ), 2 correlation measures). Of these 720 correlations, 691 are significant at the 0-level. Out of the 29 exceptions, 14 are correlations with the *tag* metric, which is omitted in further analyses.

the full corpus. Again, the correlations with citations in the same year or in the future are comparably high. Even the *tag* measure exhibits a small average correlation, yet with a standard deviation almost as high. Again, the number of bookmarks (*post*) shows the strongest correlation; its rank correlation is comparable to correlations between post counts and arbitrary correlations on *cite* (see Section 6.3), even though we compare only the counts in one single year to citations in a single year.

To get an impression on the distribution over the individual tags (tag stems), Table 6.4 shows, for each tag, Pearson’s  $r$  and Spearman’s  $\rho$  (ordered by the latter). These values are averaged per tag over the three behavioral metrics *post*, *exp*, and *view*. We omitted the other three metrics: *tag* has shown almost no (stable) correlation in the previous experiments (Tables 6.1, 6.2, and 6.3), *exp<sub>Bib</sub>* is a submetric of *exp* with almost perfect correlation (Table 6.1), and *req* is the sum of *exp* and *view*. We can observe that Spearman’s correlations rise compared to the same average computed on the full corpus (see the last line in Table 6.4). For twelve tags, we observe average correlations larger than or equal to  $\rho = 0.3$ .

By and large, we find – as expected – higher correlations for publication subsets that are more topic-homogeneous than the full corpus. Using the tags is always possible in a bookmarking system and seems to be sufficient in order to yield solid medium correlations for the three behavioral metrics *post*, *view*, and *exp* with citations both in the present and in the future.

### 6.5.3 Prediction of Future Citations

In the last part of our analysis, we investigate whether we can detect actual predictive power in the behavioral metrics. For that purpose, we conduct the binary classification experiment described in Section 6.2.2.: For each tag, we use the subset of publications together with their usage metrics *post*, *exp*, and *view* per year. We use all publication-year pairs  $(p, y)$  where for publication  $p$  at least one of the three metrics was positive in year  $y$ .

For classifying the publication-year pairs, we test three classification algorithms (as already announced in Section 6.2.2): Random Forest and two SVMs, one with polynomial and one with a radial kernel. We compare our results to two simple baselines: *Baseline Major* is a classifier that always predicts the most frequent class from the training set. Due to the unequally sized classes, this classifier can both exceed or miss an accuracy of  $\text{acc} = 50\%$  which would be achieved by random guessing. Therefore, we also compare to the latter as *Baseline Random*.

Figure 6.3 shows the results of the three classifiers on the 30 datasets. Random Forest outperforms the random baseline on 29 datasets; for the tag stem “semant” it misses it closely with  $\text{acc} = 49.60\%$ . Baseline Major is exceeded on 28 of the 30 datasets. Following [Demšar, 2006], we conduct a sign test, which confirms that the Random Forest results are significantly better than those of the baselines ( $p$  values:  $8.68 \times 10^{-7}$  for Major and  $5.77 \times 10^{-8}$  for Random). On average Baseline Random is exceeded by 9.97 percentage points and Baseline Major by 7.13 percentage points.

Table 6.4: For each of BibSonomy’s 30 most popular tag stems, average correlations between the behavioral usage and citation counts in the future (within the next year), together with the standard deviation (sd.), ordered by their Spearman’s correlation  $\rho$ . The average values are derived from the correlations between  $cit^{+1}$  and measures *post*, *exp*, and *view* respectively. Correlations are computed over all publication-year pairs in which the publication has been used at least once (the according behavioral metric is non-zero). The last line shows the corresponding averages computed on the full corpus.

tag stem (#users)	$r$ avg. $\pm$ sd.	$\rho$ avg. $\pm$ sd.
structur (224)	0.30 $\pm$ 0.10	0.43 $\pm$ 0.03
folksonomi (278)	0.22 $\pm$ 0.09	0.38 $\pm$ 0.03
web20 (234)	0.32 $\pm$ 0.08	0.38 $\pm$ 0.05
tag (294)	0.31 $\pm$ 0.08	0.37 $\pm$ 0.05
collabor (269)	0.31 $\pm$ 0.12	0.36 $\pm$ 0.05
inform (397)	0.34 $\pm$ 0.15	0.33 $\pm$ 0.05
web (409)	0.33 $\pm$ 0.12	0.32 $\pm$ 0.04
cluster (241)	0.27 $\pm$ 0.06	0.32 $\pm$ 0.07
network (384)	0.27 $\pm$ 0.11	0.30 $\pm$ 0.04
social (354)	0.15 $\pm$ 0.10	0.30 $\pm$ 0.06
commun (339)	0.29 $\pm$ 0.10	0.30 $\pm$ 0.00
algorithm (241)	0.25 $\pm$ 0.06	0.30 $\pm$ 0.05
data (269)	0.21 $\pm$ 0.02	0.29 $\pm$ 0.05
ontolog (357)	0.26 $\pm$ 0.08	0.28 $\pm$ 0.02
search (224)	0.25 $\pm$ 0.08	0.26 $\pm$ 0.05
system (382)	0.22 $\pm$ 0.12	0.26 $\pm$ 0.04
semant (380)	0.22 $\pm$ 0.07	0.25 $\pm$ 0.04
learn (309)	0.17 $\pm$ 0.08	0.25 $\pm$ 0.07
analysi (358)	0.21 $\pm$ 0.09	0.25 $\pm$ 0.03
theori (324)	0.35 $\pm$ 0.15	0.24 $\pm$ 0.03
model (460)	0.17 $\pm$ 0.06	0.24 $\pm$ 0.04
knowledg (261)	0.18 $\pm$ 0.02	0.23 $\pm$ 0.02
languag (219)	0.17 $\pm$ 0.05	0.22 $\pm$ 0.05
evalu (285)	0.25 $\pm$ 0.07	0.21 $\pm$ 0.02
internet (226)	0.18 $\pm$ 0.06	0.21 $\pm$ 0.05
softwar (277)	0.28 $\pm$ 0.10	0.21 $\pm$ 0.01
comput (317)	0.51 $\pm$ 0.08	0.20 $\pm$ 0.07
design (289)	0.48 $\pm$ 0.09	0.20 $\pm$ 0.06
process (254)	0.17 $\pm$ 0.06	0.18 $\pm$ 0.04
manag (298)	0.11 $\pm$ 0.05	0.17 $\pm$ 0.04
full corpus	0.15 $\pm$ 0.01	0.15 $\pm$ 0.01

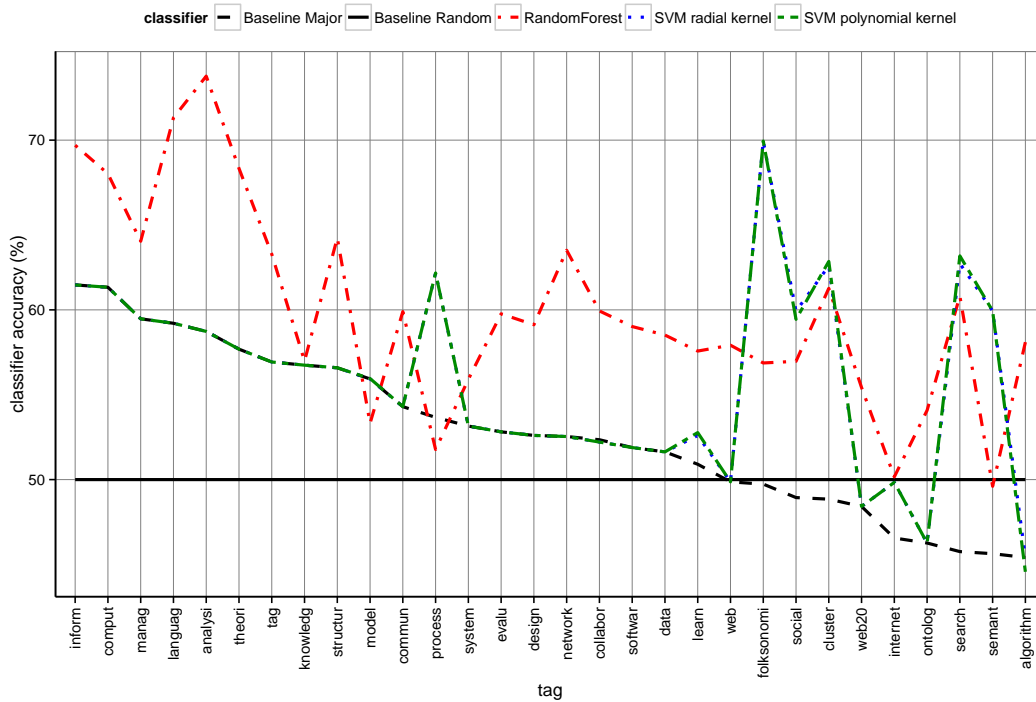


Figure 6.3: Prediction accuracy for the 30 tag-induced publication subsets for three classifiers and two baselines. Note that the diagrams for the two SVMs are almost indiscernible as they often yield identical results. The subsets are ordered by the share of entities in the test set that belong to the class that is most frequent in the training set.

The Wilcoxon signed-rank test confirms that the differences are significant evidence to reject the hypothesis, that classifier and baseline would be equally good predictors ( $p$  values:  $1.86 \times 10^{-8}$  for Major and  $5.59 \times 10^{-9}$  for Random).

On average, the two SVMs are less successful than Random Forest, although they occasionally yield better results (e.g., for “folksonomi”). Yet, they are still better than the random baseline in 25 of the 30 cases, with  $p$  values below  $5.00 \times 10^{-4}$  for both the sign test and the Wilcoxon signed-rank test. The average improvements are 5.82 percentage points (polynomial kernel) and 5.83 percentage points (radial kernel). Compared to the Baseline Major, there are still positive average improvements (2.97 and 2.99 percentage points), however, the sign tests do not allow to reject the hypothesis that either classifier performs only equally well as that baseline. The problem here is that for many subsets – 19 out of 30 with the polynomial kernel and 21 with the radial kernel – the SVMs simply predict the same class for any publication in the test set. Thus, in these cases, the SVM yields the same predictions as the

Baseline Major, leading to a draw. For some tags, however, the SVM with either kernel outperforms the baselines and also the Random Forest. This observation gives rise to the assumption that a more careful selection of the SVM's kernel might improve the prediction quality in the other cases, which is, however, beyond the scope of this thesis.

## 6.6 Conclusion

With the analyses in Section 6.5, we can answer the three research questions from the introduction. Attending to RQ1, regarding correlations on the full BibSonomy dataset, we observed small, yet noticeable correlations between citations and posting, viewing, and exporting publications – for citations in general, but also for citations occurring in the near future. We conclude that the community of all users is indeed biased towards using publications that are relevant already and also towards using publications that will become relevant soon (in terms of receiving citations). In fact, these users might well belong to those authors who cite these publications in their upcoming work. Solid correlations could be observed mainly with the metrics counting posts, exports, and views of a publication. This set of metrics yields diverse impressions on the impact of a publication, yet all three metrics exhibit medium correlations with the number of future citations.

To answer the second research question (RQ2), we grouped publications by their tags. We saw increased correlations between the usage of publications within such a subset and their citations. Altmetrics can therefore rely on the tagging feature to create subsets with stronger correlations. Tagging is inherent in a bookmarking system and no further effort, like obtaining information about publications' venues or those venues' popularity or their discipline, is required.

Finally, regarding RQ3 on the predictability of citation impact, we found small, yet significant predictive power over future citations within the three usage features we investigated (bookmarks, views, and exports). We saw that the Random Forest algorithm was able to produce significantly better predictions than the baselines. The observation of small predictive power does not justify the application in tools to actually predict citations. However, the experiment serves as a proof of concept that altmetrics like these three measures can indeed provide indicators for future citations.

Due to the limitations mentioned in Section 6.2.3, our results must be interpreted with care. Since our analysis is conducted from the point of view of BibSonomy, it is limited to this particular system and covers only a small part of the body of all scholarly articles. Yet, for the idea of exploiting usage metrics within BibSonomy, for example, for ranking or recommendations, our results are promising: They show that several usage metrics have the potential to serve as indicators for impact.

For the vision of altmetrics our results are encouraging. Our metrics are correlated with citations in the future and thus indeed valuable indicators for impact that will only later be acknowledged (formally) through citations. Thus, BibSonomy is another

potential source for altmetrics. Since none of the measures is particularly strongly correlated neither with citations nor with another metric in BibSonomy, we can also conclude that they complement each other. These metrics add to the diversity of possible measures for a publication's impact and thus truly are alternative metrics (altmetrics).

### **6.6.1 Future Research**

The results in this study give rise to a variety of further studies that could shed light on the behavior of users in a social bookmarking system and its potential to generate new altmetrics. Since the correlations were small to medium, the challenge arises to construct suitable aggregates that exhibit higher correlation. These would be relevant for webmasters who want to present predictions of future relevance in their system.

Another aspect worth analyzing is that of diversity within these metrics, to distinguish between the different forms of a publication's measurable impact. More complex measures might go beyond counting a publication's usage and could include weights (e.g., a measure of the user's expertise) or aggregate features from various systems.

The prediction of future citations was demonstrated in a simple binary setting as a proof of concept. Next steps in this direction are (i) to include further metrics from more web systems to achieve higher coverage of publications and to increase the set of features per publication, (ii) to investigate deeper, which features contribute best to successful predictions, and (iii) to adapt and optimize classifiers to yield better predictions.



## **Part III**

# **Recommendations and Reviews in Social Bookmarking Systems**



## Chapter 7

### ◆ Folksonomic Recommender Evaluation



Social bookmarking systems have established themselves as an important part in today's web. In such systems, *recommender systems* support users during the posting of a resource by suggesting suitable tags or by suggesting interesting resources. Recommender algorithms have often been evaluated in offline benchmarking experiments. Yet, the particular setup of such experiments has rarely been analyzed. In particular, since the recommendation quality usually suffers from difficulties like the sparsity of the data or the cold start problem for new resources or users, datasets have often been pruned to so-called *cores* (specific subsets of the original datasets) – however without much consideration of the implications on the benchmarking results.

In this chapter, we generalize the notion of a core by introducing the new notion of a *set-core* – which is independent of any graph structure – to overcome a structural drawback in the previous constructions of cores on tagging data. We show that problems caused by some types of cores can be eliminated using set-cores. Further, we employ the use case of tag recommendation to present a thorough analysis of recommender benchmarking setups using cores. To that end, we conduct a large-scale experiment on four real-world datasets in which we analyze the influence of different cores on the evaluation of recommendation algorithms. We can show that the results of the comparison of different recommendation approaches depends on the selection of core type and level. For the benchmarking of tag recommender algorithms, our results suggest that the evaluation must be set up more carefully and should not be based on one arbitrarily chosen core type and level.

### 7.1 Introduction

Recommender systems have been integrated into a broad variety of applications. For example, recommenders for scholarly literature can support researchers in their use of publications (the third phase of the publication life cycle) by helping them find those publications that are the most relevant to them. Often, such systems have to deal with sparse data since, usually, only little or nothing is known about many users or items. Alongside work that specifically tackles this task, in the evaluation of recommender algorithms it is common to focus on a denser subset of the data (e.g., Sarwar et al.

[2001] or Jäschke et al. [2007]), that provides enough information to produce helpful recommendations. For data that can be modeled as a graph, a commonly used technique are *generalized cores* [Batagelj and Zaveršnik, 2002], which comprise a dense subgraph in which every vertex fulfills a specific constraint. For example, the degree of each node in a graph must exceed a certain threshold, the so-called *core-level*. However, the influence of these cores on the evaluation of recommendation algorithms has not been analyzed so far.

In this chapter, we investigate cores that have been used in the evaluation of folksonomic recommender systems (cf. Section 2.4.1), so-called *post-cores*. We focus on tag recommendation – the task that post-cores were originally designed for by Jäschke et al. [2007]. Although the use of cores has become quite common in tag recommender benchmarking (cf. Section 7.3), it is unclear how the choice of

- *core type* (i.e., the method to construct the core as a subset of the original dataset; we will recall and introduce various core types in Section 7.2),
- *core level* (i.e., the threshold that is imposed on some property of each data point to construct the subset; see Section 7.2, Definitions 7.1 and 7.2),

or simply the process of constructing cores influences the results of such experiments. In fact, the choice of these setup parameters has been rather diverse in previous tag recommender benchmarking experiments (see the related work in Section 7.3 for details and for examples from the literature). Especially the core level is often set ad hoc – without a motivation for the particular choice – to such values as 2, 5, or even 100, or to dataset-dependent thresholds. In our experiments in Section 7.4, we show on real world datasets that the choice of core type and core level indeed has an impact on a benchmarking’s ranking of recommender algorithms. In fact, different experiments on different setups can lead to contradictory results. Thus, much like the choice of the evaluation metric or the sampling of training and test data, the core type and level are important aspects of an experimental setup. During the evaluation of different recommendation algorithms or during parameter optimization of such algorithms, it is therefore worthwhile to experiment with several cores and also to use the raw datasets (the unrestricted datasets). Moreover, the choice of particular core-levels should be motivated by the use case and comparisons of results from different experiments must consider the different core setups in each experiment.

While the previously used cores do yield denser graphs, they also come with the unpleasant property of *diminishing posts*: A post of the raw dataset – consisting of a user, a resource, and several tags – might still occur in the core but with fewer tags. Thus, the core construction not only reduces the number of posts in the dataset, but modifies the posts themselves. For such “diminished” posts the recommendation problem becomes more difficult as fewer tags will be considered good recommendations.

We show that cores of real-world datasets indeed contain many such diminished posts and that different recommender algorithms often yield lower quality scores on such posts than on those that still remain intact with all their tags in the core. To

overcome this structural problem we first generalize the notion of *generalized cores* from Batagelj and Zaveršnik [2002] even further to yield *set-cores* (Section 7.2). In contrast to generalized cores, these do not require a graph structure and can be applied to any kind of dataset in which the entities have some measurable property. We show that set-cores have similar properties as generalized cores, describe a construction algorithm, and prove its correctness. We then construct a new kind of core – a set-core – for social bookmarking systems which guarantees to leave all remaining posts intact (undiminished).

**Research Questions.** Before we conduct our own recommender benchmark experiments in Chapter 8, in this chapter, we investigate and question the use of cores in benchmarks of folksonomic recommender algorithms. In that respect, our research questions are:

- (RQ1) How can the notion of cores be extended to yield more flexible cores that avoid diminished posts?
- (RQ2) What influence does the choice of cores have on the result of recommender benchmarks in tagging systems?
- (RQ3) Which pitfalls do result from using *LeavePostOut* (see Section 2.4.2) with any kind of core?

**Contributions.** The contributions of this chapter are fourfold:

1. We generalize the notion of generalized cores to *set-cores* and introduce new cores for tagging data of social bookmarking systems to eliminate the particular anomaly of diminished posts in previously used cores.
2. We present a thorough investigation of the influence of cores on the results of tag recommender benchmarking experiments and confirm that different choices of core type and level can indeed yield different results.
3. We discuss potential pitfalls of the use of cores in recommender evaluation.
4. We provide recommendations for the use of cores in future recommender benchmarking experiments.

**Limitations.** Our experimental evaluation is limited to the tag recommendation scenario in tagging systems. However, the scenario suffices to demonstrate that the results of a recommender benchmarking can depend on the particular setup. Through using data from three different real-world systems, we can safely assume that we do not merely observe some peculiarity of a single dataset.

Although deficiencies of previous folksonomy cores have been the main motivation to introduce set-cores, their definition is a generalization that is independent of folksonomies and, in fact, independent of any graph structure. We mention examples for cores of other data structures in Section 7.2.3.

**Structure.** In the next section, we devise our notion of set-cores. In Section 7.3, we discuss how the experimental setup and evaluation using cores has been handled in previous work on tag recommendations. We then choose a common setup and describe in Section 7.4 several experiments on four publicly available real-world datasets to investigate the influence of cores on the results of recommender systems benchmarking. We discuss the results of these experiments in Section 7.5, where we show how different cores can lead to contradictory results in the comparison of algorithms. We also point to a peculiarity that arises from using any type of core in that setup. Section 7.6 concludes the chapter with lessons learned.

The results in this chapter have previously been published partially in [Doerfel and Jäschke, 2013] and fully in [Doerfel et al., 2016a]. They have been slightly rearranged for this thesis.

## 7.2 Cores of Graphs and Sets

Before we discuss the influence of cores within the benchmarking framework for tag recommendations, we deal with the notion of a core itself. In social bookmarking systems, the core constructions that have been used so far have the unpleasant property of diminishing posts by removing tags. In this section we present a solution to that problem by introducing post-set-cores. To accomplish that we first recall the notion of generalized cores of a graph and then extend it to arbitrary sets by introducing set-cores (Section 7.2.1). We present examples in Section 7.2.2 that illustrate different set-cores and that demonstrate some advantages of set-cores. We then discuss cores for tagging data in Section 7.2.3, where we recall the definitions of cores previously used for the evaluation of tag recommenders, and we introduce a new core construction using set-cores to overcome the issue of diminished posts.

Batagelj and Zaveršnik [2002] presented the notion of  $p$ -cores, which by itself is a generalization of the original cores introduced by Seidman [1983]. In the sequel, we refer to their construction as *graph- $p$ -cores* to better distinguish them from the new notion of *set- $P$ -cores*, which we introduce later in this section. The idea of graph- $p$ -cores is to restrict a given graph by removing all nodes for which a particular quantity  $p$  (e.g., the vertex degree) does not exceed a given threshold  $l$  called the *core level*. The graph- $p$ -core is then the largest possible subgraph such that all its vertices have the property  $p$  (measured in that subgraph) above the threshold:

**Definition 7.1** (Graph- $p$ -Core, Main Core, cf. [Batagelj and Zaveršnik, 2002]). *Let  $G = (V, E)$  be a graph,  $l \in \mathbb{R}$ , and  $p: V \times \mathfrak{P}(V) \rightarrow \mathbb{R}: (v, W) \mapsto p(v, W)$  a vertex property function on  $G$ . A subgraph  $H = (C, E|_C)$  induced by the subset of vertices  $C \subseteq V$  is called a graph- $p$ -core at level  $l$ , iff  $l \leq p(v, C)$ , for all  $v \in C$  and  $H$  is a maximum subgraph of  $G$  with that property. A core of  $G$  with a maximum level  $l$  such that it is not empty is called the main core of  $G$ .*

An example for a property function  $p$  is the vertex degree in each subgraph – in fact, the original core definition of Seidman [1983] uses just that function instead of an arbitrary function  $p$ .

The function  $p$  is called *monotone* if and only if it fulfills

$$W_1 \subseteq W_2 \subseteq V \implies \forall v \in W_1: p(v, W_1) \leq p(v, W_2).$$

Batagelj and Zaveršnik [2002] prove that for every monotone vertex property function  $p$ , a graph- $p$ -core is uniquely determined at each level  $l$  and that it can be computed by iteratively removing vertices  $v$  from the vertex set  $W$  (starting with  $W := V$ ) that do not fulfill the requirement  $l \leq p(v, W)$ . The monotonicity assumption is a mild requirement, as typical vertex property functions (like the degree) naturally fulfill it.

Note, that in Definition 7.1, the value  $p(v, W)$  of the function  $p$  is dependent both on the vertex  $v$  and on the vertex set  $W$  of the subgraph that it is evaluated on. For example, the degree of a vertex in a subgraph of  $G$  can be smaller than its degree in  $G$ . Alternatively, the function  $p$  in Definition 7.1 can be replaced by a set of functions

$$\{p_W: W \rightarrow \mathbb{R} \mid (W, E|_W) \text{ is a subgraph of } G\}.$$

A drawback of Definition 7.1 is that it is not possible to model simultaneous restrictions of several properties (e.g., that a vertex has at least in-degree  $l$  and at least out-degree  $m$ ). This can be desirable, for example, when vertices are entities in a tagging system and we want to require that each user has at least  $l$  posts, each resource occurs in at least  $m$  posts, and each tag has been used at least  $n$  times. For the case of two thresholds on a bipartite graph a solution was offered by Ahmed et al. [2007], who introduced *graph- $(p, q)$ -cores*. The *set- $P$ -core*, which we introduce next, allows us to enforce different thresholds in a more general way by using an arbitrary partially ordered set<sup>1</sup> as range for property functions, rather than only the real numbers like in Definition 7.1.

Another drawback of Definition 7.1 is the dependence on a graph structure. While this is quite universal already – as almost any kind of data can be modeled as a graph – it is not always particularly intuitive to construct a graph such that a graph- $p$ -core can be constructed. In contrast, set- $P$ -cores can be constructed on arbitrary sets.

### 7.2.1 Generalization

In the following, we present the definition of a *set- $P$ -core*, prove its uniqueness and describe a construction. A set- $P$ -core can be constructed on some arbitrary set  $S$  where for each element of  $S$  some property can be measured. Again, a threshold  $l$  is imposed to restrict the set to such elements where that property is above the threshold.

<sup>1</sup>A set  $L$  together with a binary relation  $\leq \subseteq L \times L$  is a *partially ordered set*, iff  $\leq$  is reflexive, transitive, and anti-symmetric. For two elements  $a$  and  $b$  of the partially ordered set  $L$ , we denote by  $a \not\leq b$  that  $a$  is not smaller than or equal to  $b$ , and thus that either  $a$  is larger than  $b$  or that  $a$  and  $b$  are incomparable.

In contrast to graph- $p$ -cores, the level  $l$  need not necessarily be a real number but must simply belong to some partially ordered set  $L$ , like for example, the space  $\mathbb{R}^n$ . Thus the properties are also no longer required to yield a single number, allowing to enforce multiple property restrictions simultaneously. Given a set  $S$ , the set- $P$ -core is the largest subset, such that for each element of  $S$  the chosen property functions  $P$  yield a value that is larger than a fix level  $l$ . This is stated more formally in the next definition:

**Definition 7.2** (Set- $P$ -Core,  $l$ -property). *Let  $S$  be a set,  $L$  a partially ordered set with the order relation  $\leq$ ,  $l \in L$ , and  $P$  a set of property functions  $p_{\tilde{S}}: \tilde{S} \rightarrow L$  with  $s \mapsto p_{\tilde{S}}(s)$  for each subset  $\tilde{S} \subseteq S$ . A subset  $C \subseteq S$  is said to have the  $l$ -property with respect to  $P$ , if it satisfies the condition  $l \leq p_C(c)$  for all  $c \in C$ . The subset  $C$  is called set- $P$ -core at level  $l$  of  $S$ , iff it is a maximum subset of  $S$  with the  $l$ -property.*

We simply say that a subset of  $S$  has the  $l$ -property, if the choice of  $P$  is clear from the context. Note, that in contrast to the generalized graph- $p$ -cores in [Batagelj and Zaveršnik, 2002], Definition 7.2 does neither require any kind of graph structure, nor a linearly ordered set (like the real numbers for graph- $p$ -cores).

It is easy to see that graph- $p$ -cores are special set- $P$ -cores: In the notions of Definitions 7.1 and 7.2, we set  $S := V$  (the vertex set of  $G$ ),  $L := \mathbb{R}$  and use the set of  $p$ -functions as  $P$  such that  $p_{\tilde{S}}(s) := p(s, E|_{\tilde{S}})$ . A trivial observation is that the empty set  $\emptyset \subseteq S$  has the  $l$ -property with respect to  $P$  for any  $P$  and  $l \in L$ , and thus any set  $S$  has at least one subset with the  $l$ -property.

Similar to graph- $p$ -cores, the unique existence of the set- $P$ -core is guaranteed as long as the property functions in  $P$  satisfy a mild monotonicity requirement: in each subset of the original set, for each element, the property measured by the according map in  $P$  is lower than or equal to the according value measured in the original set. Furthermore, set- $P$ -cores are nested in the sense that increasing the level  $l$  yields a smaller core. These properties are formalized and proven in the following theorem:

**Theorem 7.3.** *Given  $S$ ,  $L$ , and  $P$  as in Definition 7.2. If the functions in  $P$  are monotone in the sense that*

$$\tilde{S}_1 \subseteq \tilde{S}_2 \subseteq S \implies \forall s \in \tilde{S}_1: p_{\tilde{S}_1}(s) \leq p_{\tilde{S}_2}(s)$$

*holds, then for  $l, l_1, l_2 \in L$  hold:*

1. *The union of subsets of  $S$  with the  $l$ -property has the  $l$ -property.*
2. *There exists exactly one set- $P$ -core at  $l$ .*
3. *The set- $P$ -cores are nested, meaning, if  $l_1 \leq l_2$ , then the set- $P$ -core at  $l_2$  is contained in the set- $P$ -core at  $l_1$ .*

*Proof.* We start with the first property: Let  $I$  be an index set and  $\tilde{S}_i$  ( $i \in I$ ) be subsets of  $S$  with the  $l$ -property and  $U$  their union. For  $s \in U$ , there is some  $i \in I$



**ALGORITHM 1:** Naive set- $P$ -core construction.**Input:** Dataset  $S$ , level  $l$ , monotone set of functions  $P$ .**Output:** The set- $P$ -core  $C$  of  $S$  at level  $l$ . $C := S$ ;**while**  $\exists s \in C$  such that  $l \not\leq p_C(s)$  **do** $C := C \setminus \{s\}$ ;**end**

such that  $s \in \tilde{S}_i$ . By monotonicity of  $P$  we have  $l \leq p_{\tilde{S}_i}(s) \leq p_U(s)$  and thus  $U$  has the  $l$ -property. The second property follows directly from the first, with the set- $P$ -core being the union of all subsets of  $S$  having the  $l$ -property (and thus obviously being maximal). For the third property, let  $C_1$  and  $C_2$  be the respective set- $P$ -cores at  $l_1$  and  $l_2$ . Then  $l_1 \leq l_2 \leq p_{C_2}(s) \leq p_{(C_1 \cup C_2)}(s)$ , for  $s \in C_2$ . Since by definition  $l_1 \leq p_{C_1}(s) \leq p_{(C_1 \cup C_2)}(s)$  holds for  $s \in C_1$ , it follows that  $(C_1 \cup C_2)$  has the  $l_1$  property. By the maximality of  $C_1$  as core at level  $l_1$  follows  $C_1 = (C_1 \cup C_2)$  and thus  $C_2 \subseteq C_1$ .  $\square$

We have now established a generalized notion of cores and can reuse the simple construction algorithm from Batagelj and Zaveršnik [2011] for such a set- $P$ -core, given a finite set  $S$  (see Algorithm 1). The set- $P$ -core can always be constructed simply by removing one element violating the  $l$ -property after another until the remaining set of elements satisfies the  $l$ -property. Note however, that it does not suffice to test each element only once, as the value of the property function depends both on the element and the (remaining) subset. Thus, through the removal of other elements, the value of the property function might have decreased (in comparison to the same value before that removal) and thus might be no longer larger than the threshold  $l$ . We prove the applicability of Algorithm 1 in the following theorem:

**Theorem 7.4.** *Given  $S$ ,  $L$ ,  $l \in L$ , and  $P$  as in Definition 7.2 with  $P$  being a set of monotone functions. If  $S$  is finite, then Algorithm 1 returns the set- $P$ -core at  $l$  of  $S$ .*

*Proof.* Let  $D$  be the algorithm's result and let  $C$  be the set- $P$ -core of  $S$  at  $l$ . The unique existence of  $C$  is already guaranteed by Theorem 7.3. From the algorithm it is clear that  $D$  has the  $l$ -property and is therefore a subset of  $C$ . Let further  $s_1, s_2, \dots, s_n$  be the elements of  $S \setminus D$  in the order of their deletion by the algorithm. Assume  $D \subset C$ . Then we can select an index  $i$  with  $1 \leq i \leq n$  such that for all  $j$  with  $1 \leq j < i$ ,  $s_j$  is in  $S \setminus C$  but  $s_i$  is in  $C$ . We set  $\tilde{S}_i := S \setminus \{s_1, s_2, \dots, s_{i-1}\}$  and yield  $l \not\leq p_{\tilde{S}_i}(s_i)$ , since  $s_i$  was removed in step  $i$  of the algorithm. From the selection of  $i$  follows  $C \subseteq \tilde{S}_i$  and thus by monotonicity of  $P$  we have  $p_C(s_i) \leq p_{\tilde{S}_i}(s_i)$  and therefore  $l \not\leq p_C(s_i)$ . This is a contradiction to the  $l$ -property of  $C$ . We have thus established  $D = C$  and conclude that the algorithm's result is the set- $P$ -core at  $l$  of  $S$ .  $\square$

Table 7.1: A toy example for different cores in a user-item co-occurrence setting. The set  $U$  contains six users  $u_1, u_2, \dots, u_6$  and the set  $I$  contains six items  $1, 2, \dots, 6$ . The first two columns show the full dataset (column one shows the users and the second column their co-occurring items). Each further column  $A_1, A_2, B, \dots, F$  shows a different restriction of that dataset. The functions to create these subsets are described in Section 7.2.2.

$U$	dataset	$A_1$	$A_2$	$B$	$C$	$D$	$E$	$F$
$u_1$	1 2 3 4	1 2 3 4	1 2 3 4	1 2 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4
$u_2$	1 2 4	1 2 4	1 2 4	1 2 4	1 2 4	1 2 4	1 2 4	1 2 4
$u_3$	1 3 4	1 3 4	1 4	-	1 3 4	1 3 4	1 3 4	1 3 4
$u_4$	3 5 6	3 5 6	-	-	-	3 5 6	3 5	3 5
$u_5$	2 5	2	2	-	-	-	2 5	-
$u_6$	1 2 4	1 2 4	1 2 4	1 2 4	1 2 4	1 2 4	1 2 4	1 2 4

## 7.2.2 Examples

Our generalization allows us to transfer the concept of a core to arbitrary algebraic structures without constructing graphs. Although it is almost always possible to model a given dataset as a graph, it is not always convenient to impose a graph model. It is especially unpleasant when data is already modeled as a graph (like in the case of social bookmarking systems in the next section) but the graph does not allow the construction of a core in the desired way and thus a new graph would have to be introduced to support it. With set-cores, this is no longer an issue.

Before we leverage set-cores to construct cores for tagging data in the next section, we discuss a very simple example where data does not have to be modeled as a graph: a core that could be used in the evaluation of item recommendation algorithms. Let  $U$  be a set of users and  $I$  a set of items. Further, let  $S \subseteq U \times I$  be the user-item-co-occurrences (i.e., the relation denoting which items a user likes). Such a setting is demonstrated with a toy example in the first two columns of Table 7.1, where six users co-occur with (e.g., have expressed interest or have bought) six items in 18 user-item-co-occurrences.

Now, let  $P$  be a set of maps  $p_{\tilde{S}}$  (for every  $\tilde{S} \subseteq S$ ) with

$$p_{\tilde{S}} : \tilde{S} \rightarrow \mathbb{N} : (u, i) \mapsto \max \left( \left| \{j \in I \mid (u, j) \in \tilde{S}\} \right|, \left| \{v \in U \mid (v, i) \in \tilde{S}\} \right| \right). \quad (7.1)$$

For a given level  $l \in \mathbb{N}$ , the set- $P$ -core at  $l$  then contains all user-item-co-occurrences from  $S$  such that its user occurs with at least  $l$  items or its item occurs with at least  $l$  users. Thus, at least one entity of each user-item-pair is frequent in the dataset. In the toy example in Table 7.1, for each user-item-pair  $(u, i)$ , the maximum of its user and item frequency is larger than or equal to two. Therefore, the set- $P$ -core at level  $l = 2$  is the full dataset.

For  $l = 3$  we obtain the dataset denoted by  $A_1$  (third column) in which only the user-item-pair  $(u_5, 5)$  has been removed, as both user  $u_5$  and item 5 occur in only two user-item-pairs. In this first example, we can observe that the resulting core actually has a lower density<sup>2</sup> than the original dataset ( $\frac{17}{36}$  versus originally  $\frac{18}{36}$ ), since through the removal of the pair  $(u_5, 5)$  neither a user nor an item have been removed completely from the dataset. This might reduce the computational complexity in an item recommender scenario (for algorithms that depend on the number of pairs) but usually, artificially introducing sparseness is not desirable. The next examples will show cores where the density rises. Furthermore, we will see in Section 7.4.1 that our core constructions yield an increase in density on all four real world datasets.

Increasing the level to  $l = 4$  yields the core denoted by  $A_2$  in Table 7.1. Here, all pairs are removed where user and item both occur in less than four pairs. These are all pairs containing user  $u_4$  and all pairs containing items 5 or 6. Thus, these three entities can be removed from the dataset completely. In comparison to the core for  $l = 3$ , we now indeed yield a dataset with higher density than the original dataset ( $\frac{13}{20}$  versus originally  $\frac{18}{36}$ ).  $A_2$  is the main set- $P$ -core, as for  $l = 5$  the core vanishes since no user nor item occurs in more than four pairs.

Using the minimum instead of the maximum in the definition of  $p_{\tilde{S}}$  in Equation 7.1, results in a core containing user-item-co-occurrences where both user *and* item are frequent, as here the smaller of the two frequencies – and thus both frequencies – must exceed the threshold  $l$ . In the toy example in Table 7.1, the set- $P$ -core for  $l = 3$  is denoted by  $B$ . It can be constructed using Algorithm 1 by first removing the co-occurrences  $(u_4, 5)$ ,  $(u_5, 5)$ , and  $(u_4, 6)$ , since items 5 and 6 both are not frequent. The pair  $(u_5, 2)$  is removed since user  $u_5$  is not frequent. Then the remaining co-occurrence of user  $u_4$  –  $(u_4, 3)$  – is removed, since after the elimination of  $(u_4, 5)$  and  $(u_4, 6)$ ,  $u_4$  has become infrequent. Then all co-occurrences with item 3 and finally those of user  $u_3$  must be removed. In the example, the set  $B$  is the main set- $P$ -core since for level  $l = 4$  all user-item-pairs would be removed from the dataset.

An example for a core, where different thresholds can be imposed on users and items, results from the maps:

$$p_{\tilde{S}}: \tilde{S} \rightarrow \mathbb{N}^2: (u, i) \mapsto \left( \left| \left\{ j \in I \mid (u, j) \in \tilde{S} \right\} \right|, \left| \left\{ v \in U \mid (v, i) \in \tilde{S} \right\} \right| \right) \quad (7.2)$$

together with a level  $l := (l_u, l_i) \in \mathbb{N}^2$ . This setting yields a core where each user occurs with at least  $l_u$  items and each item with at least  $l_i$  users. Thus, we have made use of two thresholds at the same time, which could not have been modeled with graph-cores. In the toy example in Table 7.1, dataset  $C$  shows the  $(3, 2)$ -set-core. In contrast to the previous result  $B$  – where user and item both had to occur three times in the dataset – item 3 still has co-occurrences in the dataset. Note, that setting two thresholds  $l_u$  and  $l_i$  at the same time is not the same as first setting one threshold  $l_u$  on the users, then setting one threshold  $l_i$  on the items and taking the intersection of

<sup>2</sup>Density in this dataset is computed as the number of actual user-item-co-occurrences divided by the number of possible ones.

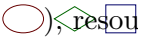



the resulting sets. The latter procedure would not necessarily yield a set where each user occurs at least  $l_u$  and each item at least  $l_i$  times. This is demonstrated in the toy example with the datasets  $D$ ,  $E$ , and  $F$  being restricted datasets with  $l_u = 3$  ( $D$ ) and with  $l_i = 2$  ( $E$ ), as well as their intersection ( $F$ ). We can observe that dataset  $F$  is different from  $C$  and that user  $u_4$  violates the constraint on the user frequency by having only two co-occurrences instead of the required three and item 5 does not satisfy the lower bound on the item frequency as it is part of only one co-occurrence.

In these examples, we have demonstrated the ability of set-cores to restrict datasets according to individual thresholds on different sets of entities, like in the latter example with one threshold for users and one for items. We have also seen the application of combined thresholds like the first two examples, using the maximum or minimum of user-item-co-occurrences. Both aspects allow a great flexibility for the practitioner: Thresholds can be chosen individually for different entities and at the same time, combined thresholds can be imposed. For the latter, min and max are only simple examples: we could just as easily use sums, products, or other functions, as long as they comply with the monotonicity requirement in Theorem 7.3. Using the sum instead of max or min in the above example would impose a threshold for the combined popularity of user and item in a user-item-pair. Finally, it is also possible to combine the maps of the different examples in Equations (7.1) and (7.2) (and thus yield maps  $p_{\tilde{S}}: \tilde{S} \rightarrow \mathbb{N}^3$ ) to have individual requirements on users and items as well as a combined requirement for each user-item-pair.

### 7.2.3 Cores of Folksonomies

We employ the use case of social bookmarking systems and their underlying data structures called folksonomies (cf. Section 2.3) to demonstrate different types of cores. These cores are also the subject of the experimental investigation of tag recommender evaluation frameworks in the following sections. In particular, we design cores that will leave posts intact and thus respect the unit in which tag assignments are usually created and depicted in tagging systems.

**Running Example.**

users  $A, B, C$  (drawn )<sup>resources</sup> )<sup>resou</sup>, and tags 1, 2, 3 (  ) are connected by thirteen tag assignments in seven posts (numbers 1-7). The hyperedges that represent the tag assignments are visualized by  which are connected to the three vertices of each hyperedge. The number next to each circle depicts the number of the post the tag assignment belongs to. The differently colored areas that enclose parts of the graph depict the various types of cores that are explained in the sequel.

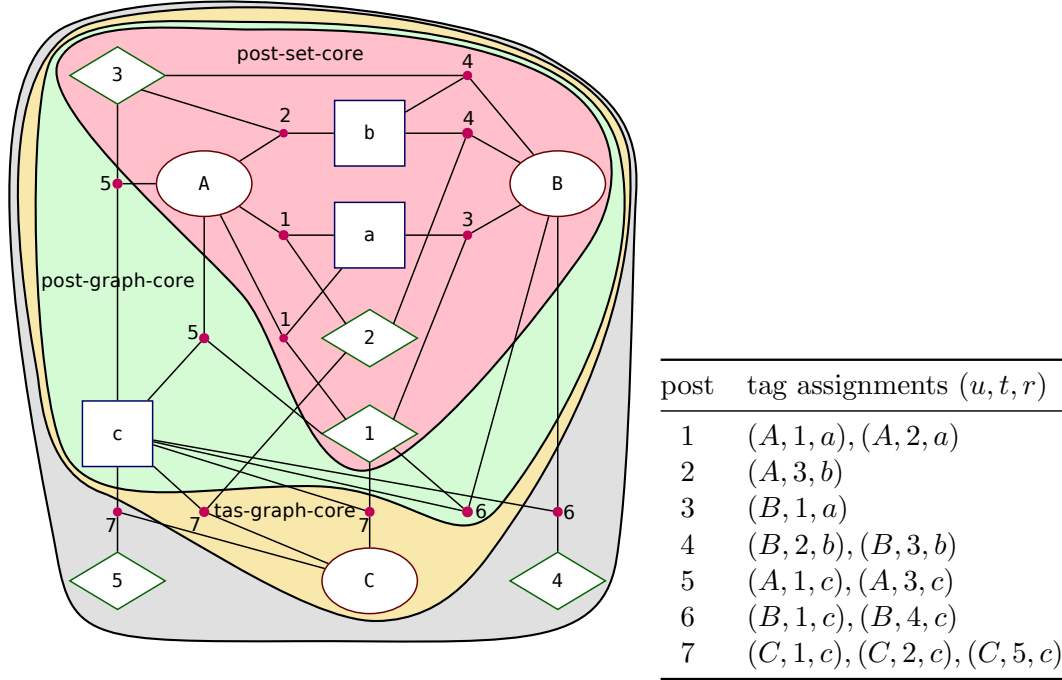


Figure 7.1: A folksonomy toy example with a tas-graph-core, post-graph-core, and post-set-core. The table on the right lists the tag assignments that belong to each post.

### The Tas-Graph-Core of a Folksonomy

We regard a folksonomy  $\mathbb{F} = (U, T, R, Y)$  as a hypergraph  $G = (V, E) := (U \cup T \cup R, Y)$ . Together with the level  $l \in \mathbb{N}$  and the vertex property function

$$p: V \times \mathfrak{P}(V) \rightarrow \mathbb{N}: (v, W) \mapsto \begin{cases} |(\{v\} \times T \times R) \cap E|_W & \text{if } v \in U \\ |(U \times \{v\} \times R) \cap E|_W & \text{if } v \in T \\ |(U \times T \times \{v\}) \cap E|_W & \text{if } v \in R, \end{cases} \quad (7.3)$$

that assigns to every  $W \subseteq V$  and every  $v \in W$  the number of tag assignments (tas) that  $v$  is part of in  $W$ , we get the *tas-graph-core at level  $l$  of the folksonomy  $\mathbb{F}$* . It has the property that every user, tag, and resource is part of at least  $l$  tag assignments. Note, that for a tag to be part of a tas-graph-core at level  $l$ , it must have been used in at least  $l$  posts, while for a user (resource) it is sufficient to annotate (be part of) only a single post with at least  $l$  tags (cf. Jäschke et al. [2008]).

**Running Example.** Figure 7.1 shows the *tas-graph-core* at level 2, in which every entity belongs to at least two tag assignments. The tag assignments  $(B, 4, c)$  from post 6 and  $(C, 5, c)$  from post 7 are lost because the tags 4 and 5 belong each only to the one corresponding tag assignment. Note, that the *tas-graph-core* does *not* have

level 3, since the tag assignment  $(C, 5, c)$  does not belong to the tas-graph-core and thus user  $C$  occurs in only two tag assignments.

### The Post-Graph-Core of a Folksonomy

To circumvent the aforementioned problem, Jäschke et al. [2007] (and more formally [Jäschke et al., 2008]) defined another core, here called *post-graph-core* to distinguish it from the other types of cores. It uses another vertex property than the tas-graph-core: For  $v \in W \subseteq V$  it counts for every entity of the folksonomy the number of *posts* (instead of tag assignments, as before) this entity is part of. Let  $T_{ur}$  denote the set of tags, which user  $u$  assigned to resource  $r$  in  $\mathbb{F} : T_{ur} := \{t \in T \mid (u, t, r) \in Y\}$ . Then, the post-graph-core is constructed with the vertex property:

$$P: V \times \mathfrak{P}(V) \rightarrow \mathbb{N}: (v, W) \mapsto \begin{cases} |\{(v, T_{vr} \cap W, r) \mid r \in R \cap W, T_{vr} \cap W \neq \emptyset\}| & \text{if } v \in U \\ |\{(u, v, r) \in E \mid_W\}| & \text{if } v \in T \\ |\{(u, T_{uv} \cap W, v) \mid u \in U \cap W, T_{uv} \cap W \neq \emptyset\}| & \text{if } v \in R \end{cases}$$

This definition intuitively violates the symmetry of the ternary structure of a folksonomy: In contrast to the previous core, the value of the property function  $P$  for tags ( $v \in T$ ) is no longer defined analogously to the values for users and resources. This is because one post always contains exactly one user and exactly one resource but can have more than one tag. However, the post-graph-core more closely resembles the view of a folksonomy as ‘a collection of posts’ that collaborative tagging systems typically provide. The post-graph-core at level  $l$  has the property that each user, tag, and resource occurs in at least  $l$  posts. In Section 7.3, we list various examples for the frequent use of post-graph-cores in the evaluation of recommender systems for folksonomies.

**Running Example.** In the example in Figure 7.1 we can see that in the *post-graph-core* at level 2, every entity belongs to at least 2 posts. Due to the removal of the user  $C$  (which only belongs to post 7), all tag assignments from post 7 (and thus the post itself) are removed. Similarly, tag 4 is removed as it belongs only to post 6.

### Diminished Posts in Tas-Graph-Cores and Post-Graph-Cores

In the previous two constructions, the core is computed by removing single tag assignments. Thus, from one post, several tag assignments can be removed, while others (of the same post) remain in the core. This rather unfortunate behavior is illustrated in Table 7.2 using a post from the BibSonomy *book* dataset which we use and describe in Section 7.4.1. The post (that is also shown in Figure 2.2) consists of five tag assignments in the original dataset. By restricting the data to a tas-graph-core or a post-graph-core, some of these tag assignments are omitted and the post is diminished. In the tas-graph-core at level 2, first the two rare tags “requetes” and “webmetrics” vanish from the post. At level 3 and also in the post-graph-cores, also

Table 7.2: An example post from the *book* dataset (cf. Section 7.4.1) that is diminished by the construction of cores. In the post – stored by a user with ID 1015 – the resource is a bookmark to the URL <http://www.google.com/trends> and the post was annotated with the five tags “statistics”, “trends”, “comparateur”, “requetes”, and “webmetrics” (see also Fig. 2.2). Through the core constructions, some tags are removed from the data, while others remain (column “tags”) leaving the post diminished in the respective core. Eventually, for tas-graph-cores at levels  $l \geq 13$  and post-graph-cores at levels  $l \geq 6$  the post vanishes completely from the core.

core	tags
full dataset	statistics, trends, comparateur, requetes, webmetrics
tas-graph-core, $l = 2$	statistics, trends, comparateur
tas-graph-core, $3 \leq l \leq 12$	statistics, trends
post-graph-core, $2 \leq l \leq 5$	statistics, trends

the tag assignment with the tag “comparateur” is removed. The tags “statistics” and “trends” are well connected with other folksonomy entities through tag assignments in other posts. Thus, they remain in the cores for several levels until the complete post vanishes from the dataset at levels higher than 12 for tas-graph-cores and levels higher than 5 for post-graph-cores.

**Running Example.** In our running example, we can also observe a diminished post. In the constructions of both the tas-graph-core and the post-graph-core, post 6 is diminished: tag assignment  $(B, 1, c)$  still belongs to the core, while tag assignment  $(B, 4, c)$  is removed. Thus post 6 has now only one tag, instead of the original two.

The use of set-cores now allows us to overcome the phenomenon of diminished posts by regarding posts as atomic entities. We call this new construction the *post-set-core* of a folksonomy:

### The Post-Set-Core of a Folksonomy

Let  $S$  be the set of all posts in  $\mathbb{F}$ , and for some subset  $\tilde{S} \subseteq S$  of posts, let  $\tilde{S}_u, \tilde{S}_t, \tilde{S}_r$  be the sets of posts in  $\tilde{S}$ , that a user  $u$ , a tag  $t$ , or a resource  $r$  occurs in, respectively. Note, that these can be empty sets, if the according entity of  $\mathbb{F}$  does not occur in any post contained in  $\tilde{S}$ . Now, the functions

$$p_{\tilde{S}}: \tilde{S} \rightarrow \mathbb{N}^3: (u, T_{ur}, r) \mapsto \left( |\tilde{S}_u|, \min_{t \in T_{ur}} |\tilde{S}_t|, |\tilde{S}_r| \right)$$

are monotone in the way required by Theorem 7.3 (where  $\mathbb{N}^3$  is partially ordered as usual by  $(a_1, b_1, c_1) \leq (a_2, b_2, c_2) \iff a_1 \leq a_2, b_1 \leq b_2, c_1 \leq c_2$ ). The monotonicity simply follows from the observation that by shrinking  $\tilde{S}$  the sets  $\tilde{S}_u, \tilde{S}_t$  and  $\tilde{S}_r$  can lose but never gain cardinality.

The functions assign to each post a triple: the number of posts that the user, the rarest tag (counted in  $\tilde{S}$ ) and the resource of the post is part of within the subset  $\tilde{S}$ . For any vector  $l \in \mathbb{N}^3$ , we can now construct a *post-set-core* as a set- $P$ -core at  $l$ . In particular, this notion of core allows us to select three different thresholds  $(l_u, l_t, l_r) \in \mathbb{N}^3$  for the number of occurrences of users, tags, and resources, respectively. The following examples illustrate use cases for choosing different thresholds:

- When one goal of a tag recommender is to consolidate the tag vocabulary of the system, a large threshold  $l_t$  ensures that only frequently used tags remain in the dataset for evaluation. The thresholds  $l_u$  and  $l_r$  can remain low.
- If the cold-start problem for users and resources shall be neglected in the evaluation, high values for  $l_u$  and/or  $l_r$  can be selected while  $l_t$  can be low.

For the sake of simplicity, we say that a post-set-core is at level  $l$  when all three thresholds are equal to  $l$ .

With the introduction of post-set-cores, we have answered our first research question (RQ1): post-set-cores do not contain diminished posts and they allow individual thresholds on users, tags, and resources, thus they allow flexibility and modeling close to the actual use case.

**Running Example.** The *post-set-core* at level  $(2, 2, 2)$  is shown in Figure 7.1, where every user, tag, and resource of the four posts 1, 2, 3, and 4 belongs to at least two of these four posts. The example also illustrates an important property of the post-set-core construction: None of the remaining posts is diminished, all remaining posts are complete in the sense that they still contain all the tags they have in the original dataset, as each post as a whole is treated as an atomic part of the dataset. This is neither the case for the tas-graph-core (e.g., post 7 loses tag 5) nor in the post-graph-core (post 6 loses tag 4) since here the posts are modeled as collections of tag assignments and tag assignments are removed individually during the core construction.

The example in Figure 7.1 also illustrates the property that a tas-graph-core always contains the post-graph-core at the same level, and the latter contains all posts of the post-set-core at that level. This property follows directly from the core construction and is formalized in the following lemma.

**Lemma 7.5.** *Given a folksonomy  $\mathbb{F}$  and a level  $l \in \mathbb{N}$ .*

1. *Each user  $u$ , tag  $t$ , and resource  $r$ , as well as each tag assignment  $(u, t, r)$  of the post-graph-core at level  $l$  is contained in the tas-graph-core at level  $l$ .*
2. *For each post  $(u, T_{ur}, r)$  in the post-set-core at level  $l$ , the entities  $u$ ,  $r$ , and  $t \in T_{ur}$ , as well as all tag assignments  $(u, t, r)$  (for  $t \in T_{ur}$ ) are contained in the post-graph-core at level  $l$ .*



Finally, a trivial example for each core type is the 1-core, which is the full folksonomy itself, excluding isolated nodes (e.g., users that registered with the system but never used it and thus do not occur in a post). Tag recommender evaluation usually ignores isolated nodes and therefore the cores at level 1 are just the original evaluation datasets (in the following also referred to as the *raw data*).

### Similar Constructions on Other Data

The core constructions described for folksonomies can easily be generalized to other data structures where entities have some countable properties. For example, a tweet in the micro blogging system Twitter<sup>3</sup> consists of a user, URLs, hashtags, and several words. Much like in the case of folksonomies, we can derive countable properties for each tweet  $t$ , like the minimum number of tweets the URLs of  $t$  occur in, the minimum number of tweets that each hashtag of  $t$  occurs in, or the minimum number of tweets each word of  $t$  occurs in. Using a set-core like in 7.2.3, we can then simply impose individual thresholds on the URL, hashtag or word frequencies. Depending on the particular use case, one might, for example, set high thresholds on the hashtag and URL frequencies to select only trending topics and resources, while setting a low threshold for words. Moreover, it would be possible to combine two aspects, say URLs and hashtags, by using maps that count the number of tweets that share either the URL or a hashtag with a tweet  $t$ .

In contrast to the use of set-cores, graph-cores would require to impose a graph structure first, for example, by connecting all entities of a tweet by 4-dimensional hyperedges, where each edge connects the user to one of the hashtags, to one of the words, and to one of the URLs of a tweet. Other than with set-cores however, such graph-cores would yield “diminished tweets” (e.g., missing some infrequent words or hashtags). Furthermore, since including URLs or hashtags in a tweet is optional, the graph model would have to be able to deal with tweets that do not contain all these components (e.g., by using edges of different dimensionality).

## 7.3 Related Work

In this section, we review and discuss several examples from the literature that deal with the topics of this chapter. We start with the previous use of cores in various areas of research before we turn our attention to the evaluation of recommender systems. We discuss the well-known problem of *sparse data*, which can be tackled by focusing on the dense part of the data, for instance, by using *graph-cores*. Since the latter is often the case in the benchmarking of tag recommender algorithms, we review the state of the art in that area next, covering different approaches as well as variations in the experimental setups. Finally, we compare several previous tag recommender benchmarking studies regarding their use of cores.

---

<sup>3</sup><http://www.twitter.com/>

### 7.3.1 Graph-Cores

One widely applicable methodology to create dense subsets of graphs are the so-called *graph-cores* which were introduced by Seidman [1983]. Batagelj and Zaveršnik [2002] (and later, Batagelj and Zaveršnik [2011]) extended this work by introducing *generalized cores* – see Section 7.2 for details. Cores have previously been used to create generative models of graphs [Baur et al., 2007] or to improve the visualization of large networks [Ahmed et al., 2007]. Angelova et al. [2008] analyzed cores of various derived graphs (friendship, common entities, and similarity graphs) of a social bookmarking dataset. The number of connected components quickly drops to one, already for small core levels. In general, an increasing core level results in a decreasing average node distance and a more complex behavior of the average clustering coefficient. In [Wang and Chiu, 2008], cores of a transaction network of an online auction system were used to identify densely connected subgraphs of malicious traders in order to recommend trustworthy auction sellers. Jiang et al. [2013] compared cores of three graphs from the same social networking platform (Renren, a Chinese social network) over different levels. They found that the core size is relatively stable up to level 60 for a graph representing explicit friendships between users, but decreases much earlier for other interaction networks. Similarly, Leskovec and Horvitz [2008] investigated the decreasing core size of the Microsoft Instant Messenger communications network and Chun et al. [2008] analyzed the same property within the South Korean social network Cyworld, comparing friendship graphs and guest-book interactions.

By now, cores are an established methodology to analyze the structure and dynamics of graphs with applications as diverse as community detection [Giatsidis et al., 2011], temporal analysis of the internet topology [Alvarez-Hamelin et al., 2008], or the study of large-scale software systems [Zhang et al., 2010]. Our generalization to arbitrary sets in Section 7.2 therefore opens up new possibilities for core-based analyses on data other than graphs.

### 7.3.2 Evaluation of Recommender Systems

Research on recommender systems evaluation typically focuses on the selection of proper metrics and performance criteria, like user preference, coverage, trust, or novelty, as well as on data processing and selection methods. A good overview has been presented by Shani and Gunawardana [2011]. A fixed selection of metrics and criteria constitutes an *evaluation framework*, in which several recommender algorithms can be compared against each other in benchmarking experiments. Often however, such a framework is chosen ad-hoc and the implications of the selection rarely have been investigated. Often, several choices are valid and plausible yet can lead to contradictory results. The consequence of different choices of an evaluation metric was, for example, demonstrated by Schein et al. [2002], who compared two metrics: One metric focuses on a broad coverage of users (good recommendations for each user) while the other rewards as many good recommendations as possible independent from the distribution

over the users. In practice, the choice of the ‘best fitting’ metric is up to the operator of the actual recommender system. Said and Bellogín [2014] presented a benchmarking of recommendation frameworks. They found pronounced differences in the results of different frameworks, resulting from different experimental setups, but also from subtleties in the implementation of algorithms and metrics. Many studies comparing recommender algorithms find that (part of) their results are dataset dependent. For example, Karypis [2001] showed that in item-based collaborative filtering algorithms, varying a particular parameter (controlling the influence of popular items) yields different behavior on different datasets.

Cremonesi et al. [2010] used a movie recommendation scenario to demonstrate that different recommender algorithms respond differently to a subsampling of the test set. Their approach was to remove items from the test set that belong to the most popular (most frequent) items in the datasets. Thus, a strictly popularity-based algorithm exhibited a heavy performance decrease (compared to its score on the full test data). In contrast, other algorithms had less strongly decreased scores, such that the resulting ranking of the top performing algorithms was different to that on the full test set. The motivation for the exclusion of the most popular items in [Cremonesi et al., 2010] was to demote algorithms that tend to favor the most popular items as such items are often already known to the user and thus do not present interesting recommendations. In contrast, the core construction in the tag recommender setting is not used to filter out particularly unwanted recommendations but rather to mitigate the cold start problem. We will, however, demonstrate in Section 7.5.1 that also in this scenario, different recommenders react differently to changes of the core setup, and in Section 7.5.3, we discuss the changes in the resulting rankings of algorithms. Adomavicius and Zhang [2012] investigated for the use case of the classical recommender scenario, where users assign ratings to items, how different dataset characteristics influence the resulting scores of recommendation algorithms. Several recommender algorithms were evaluated on various subsamples of five datasets. The results show that some properties, like the size of the rating space and the data density, have a positive impact on the recommendation quality, while others have a negative impact (e.g., the standard deviation of the ratings in a dataset). Furthermore, there are properties of the rating frequency distribution that have a positive influence for some of the tested algorithms, but a negative influence on the performance of others. In our experiments in Section 7.5.1, we can similarly observe that the properties core type and core level have an influence on the benchmarking of tag recommender algorithms, and that different tag recommender algorithms react differently to different cores.

### 7.3.3 Sparse Data

In many recommendation scenarios, the sparsity of the data is a classical problem: Users use (rate, buy, tag, etc.) only a very small part of the available items. Thus, only very little is known about both users and items, making it harder to build reliable models for recommender algorithms. For example, sparse user-rating data limits

the identification of similar users and items in collaborative filtering [Sarwar et al., 2000]. The sparsity problem has been tackled either by dealing with the sparsity in particular or by focusing on the dense part of the data (e.g., Sarwar et al. [2001]). A typical approach to reduce sparsity is dimensionality reduction. For instance, Ma et al. [2008] proposed a matrix factorization approach that combines traditional rating data with social network data to reduce the sparsity of the ratings matrix. Sarwar et al. [2000] used singular value decomposition to compute user neighborhoods on dense, low-dimensional product matrices. Content-based approaches have also been used to increase the density of the ratings matrix for collaborative filtering (e.g., Melville et al. [2002]) or have been combined with collaborative methods (e.g., Popescul et al. [2001]) using a unified probabilistic framework.

Most of the approaches that focus on the dense part of the data are rather ad-hoc, usually defining some threshold for the minimal number of ratings an item or user should have. There are few theoretical considerations or experiments that investigate the implications of such thresholds on the performance of different recommender algorithms or the validity of the experiments. For example, Herlocker et al. [2004] addressed the density of datasets as one of the properties that influence recommender systems evaluation. While they empirically compared different (classes of) evaluation metrics, they do not further investigate density as a factor of the evaluation. In this work, we show how using cores (to increase the density of the data) can influence the results of a tag recommender benchmarking.

### 7.3.4 Tag Recommender Systems and their Evaluation

Since the emergence of social bookmarking, the topic of tag recommendations has raised considerable interest of researchers such that a vast body of literature exists. Here, we list a selection of these previous approaches, focusing on their various experimental setups for the comparison of different algorithms.

Mishne [2006] approached the problem of recommending tags for blog posts. The evaluation comprises a manual inspection of the recommended tags for 30 randomly selected blog posts, using precision at 10 (cf. Section 2.4.2), and an automatic comparison, using precision at 10 and recall at 10, of 6,000 randomly selected, already tagged posts. Only posts with three or more tags were considered for the automatic evaluation, and for comparing predicted and actually used tags, their string distance was used instead of exact matching, though no details about the maximal allowed string distance were given. Xu et al. [2006] identified properties of good tag recommendations, like high coverage of multiple facets, high popularity, or least-effort, and evaluated their approach qualitatively on 18 resources (URLs). Vojnović; et al. [2007] tried to imitate the learning of the true popularity ranking of tags for a given resource during the assignment of tags by users. The method was evaluated with precision over 1,200 resources, though no details about the hold-out set are given. An extensive benchmarking of collaborative filtering, the graph-based FolkRank algorithm [Hotho et al., 2006c], and simpler methods based on the usage frequency of tags was per-

formed by Jäschke et al. [2008] on three datasets from CiteULike,<sup>4</sup> Delicious,<sup>5</sup> and BibSonomy (Section 2.3.2). The evaluation was conducted using post-graph-cores in the *LeavePostOut* setup (cf. Section 2.4.2). The same setup was used by Ramezani [2011] and Seitlinger et al. [2013].

The ECML PKDD Discovery Challenges 2008 and 2009 [Hotho et al., 2008, Eisterlehner et al., 2009] both included tag recommendation tasks. In 2008, tag recommenders were tested offline and the test set comprised all posts added to BibSonomy during a period of one and a half months. The 2009 Discovery Challenge established a common evaluation protocol, against which more than 20 approaches were evaluated: on datasets from BibSonomy, posts from the most recent six months were used as test data, and the approaches were evaluated with the F1 measure<sup>6</sup> over the top five recommended tags. One task focused on graph-based recommendations and ensured that every tag, user, and resource from the test dataset were already contained in training data by using a post-graph-core at level 2. The content-based task was evaluated on the complete six months of the test data. A novelty of the challenge was the evaluation of some algorithms in an online setup where the click-rate of BibSonomy users could be measured.

Krestel et al. [2009] presented a tag recommendation algorithm based on Latent Dirichlet Allocation. The evaluation was performed per resource: Almost all posts (except up to five) for a resource were removed and the recommender then tried to predict their tags. The test data consisted of 10% of the posts and the whole experiment was repeated five times. Ma et al. [2013] proposed the algorithm *TagRank* and performed five-fold cross validation: the results were “averaged over each user, then over the final five folds” though no details on how the data was split (e.g., per user) were given.

Overall, tag recommendation<sup>7</sup> algorithms are typically evaluated on offline datasets, the posts for the test sets are selected at random, and measures like precision, recall, and F1 are used for evaluation. There is a tendency to use the *LeavePostOut* methodology, though other cross-validation procedures (*LeavePostOut* is  $|U|$ -fold cross validation where  $|U|$  is the number of users in the dataset) and other types of splits are also used.

### 7.3.5 Cores and Recommender Systems

As part of the evaluation of recommender systems, cores have first been used by Jäschke et al. [2007] to focus on the dense part of folksonomies. Experiments with different tag

<sup>4</sup><http://www.citeulike.org/>

<sup>5</sup><http://delicious.com/>

<sup>6</sup>The harmonic mean of recall and precision.

<sup>7</sup>While the here mentioned experiments all tackle tag recommendations, it is worth noting that also for resource recommendations, datasets are often restricted. For example, Bogers [2009] restricted datasets of BibSonomy, Delicious, and CiteULike such that each user had at least 20 resources and each resource occurred in posts of at least two users. Since no threshold was applied to tags, no posts are diminished in that setting.

recommenders were conducted on subsets of folksonomies, constructed as generalized cores – so-called *post-cores*, like explained in the previous section. Cores were then commonly used in the evaluation of (tag) recommendation algorithms for collaborative tagging systems, for instance, by Ramezani [2011] to compare different *PageRank* variants on cores from BibSonomy, CiteULike, and Delicious at levels 5, 5, and 20, respectively; by Krestel et al. [2009] to evaluate a tag recommendation algorithm based on Latent Dirichlet Allocation on a core at level 100 of a dataset from Delicious; by Seitlinger et al. [2013] to evaluate a category-based tag recommender on a Delicious core at level 14; by Ma et al. [2013] to evaluate a variant of topic-sensitive *PageRank* upon a tag-tag correlation graph on a Delicious core at level 9 and cores at level 5 from last.fm and Movielens; and by Nanopoulos et al. [2013] to evaluate a matrix factorization-based song recommender with the core level set such that it is equivalent to 0.001 % of the total play counts. As mentioned earlier, a post-graph-core at level 2 of a BibSonomy dataset was also used in the ECML PKDD Discovery Challenge 2009 for the graph-based task. For example, Rendle and Schmidt-Thieme [2010] used it, next to a BibSonomy post-graph-core at level 5 and a last.fm post-graph-core at level 10, to evaluate their successful tensor factorization approach to tag recommendation, which won the challenge.

This overview shows that the choice of the particular core level is very diverse and typically neither justified nor evaluated. The arguments for using cores are similar throughout these approaches and have been summarized by Ma et al. [2013]: “the size of each dataset is dramatically reduced allowing the application of recommendation techniques that would otherwise be computationally impractical, and by removing rarely occurring users, resources and tags, noise in the data can be considerably reduced.” Except for [Jäschke et al., 2008], all these works did neither question nor challenge the use of cores nor did they compare their findings on several cores or to results on the raw data. In [Jäschke et al., 2008], the results on a Delicious core at level 10 were compared to results on a dataset where only users and resources with less than two posts were removed. Recall and precision of all algorithms except the *adapted PageRank* were found to be similar. Furthermore, besides the typical lack of evaluation on the raw data, all aforementioned evaluation setups suffer from the problem of diminished posts which we described in Section 7.2.3 together with a solution by introducing post-set-cores.

### 7.3.6 Summary

As we pointed out in the previous sections, one commonly used framework for collaborative tagging systems comprises graph-cores in an offline setting where recall and precision are measured. In the experiments in the following, we do not aim at the evaluation of different properties of recommender systems nor at the presentation of a new evaluation framework. Instead, we investigate the robustness of that common evaluation framework itself and we challenge commonly used methodologies. Therefore (and to be comparable with previous works), we investigate the influence of different core

types and levels within the fixed framework for *offline* evaluation of *tag recommender systems* in *folksonomies* using *cores* in combination with the *LeavePostOut* method and the standard measures *precision*, *recall*, and *MAP*.

## 7.4 Experimental Evaluation

The main goal of the experiments in this chapter is to demonstrate how benchmarking results act over different core types and levels. In the experiments, we show that the quality of recommendations depends on (mostly increases with) the core level; that diminished posts indeed occur frequently in *tas-graph-cores* and *post-graph-cores*, and these posts influence the overall results; and that different core setups (different core types or levels) can lead to conflicting results in a benchmarking’s ranking of algorithms. Furthermore, we point to a peculiarity of using cores that arises from their use in the *LeavePostOut* evaluation scenario. To that end, we choose a fix evaluation setup for tag recommender algorithms – like it has been used frequently in previous studies – and apply it to four real world datasets. In that setup we then vary the cores and discuss the differences in the results using different metrics.

In this section, we describe the setup of our experiments to test different evaluation procedures with different cores, levels, and metrics for tag recommender algorithms. More specifically, we

- describe four datasets from three collaborative tagging systems, namely *BibSonomy*, *CiteULike*, and *Delicious*;
- explain the cleansing procedure that includes, among others, the removal of imported posts;
- show some basic properties of the datasets, like size and density for different core types and levels; and
- detail which cores, evaluation protocol, metrics, and recommender algorithms we use.

### 7.4.1 Datasets

We use four datasets from three tagging systems (for an overview see Table 7.3):

The *BibSonomy* dataset from July 1, 2012 is a regular dump of the system’s publicly available data.<sup>8</sup> *BibSonomy* supports bookmarking of both bookmarks and publication metadata, hence we split the data into two parts: *book* and *publ*. From *CiteULike*, we use the official snapshot (*cite*) from May 14, 2012.<sup>9</sup> From *Delicious*, we use a dataset (*deli*) that was obtained during July 27, 2005 and July 30, 2005, and that has previously been used by Hotho et al. [2006c].

<sup>8</sup><http://www.kde.cs.uni-kassel.de/bibsonomy/dumps/>

<sup>9</sup><http://www.citeulike.org/faq/data.adp>

Table 7.3: The sizes of the folksonomies in the four datasets.

dataset	$ U $	$ T $	$ R $	$ Y $	$ \text{posts} $
<i>publ</i>	4,777	57,639	94,427	397,081	109,984
<i>book</i>	4,959	80,603	231,907	1,032,037	268,589
<i>deli</i>	75,071	397,028	2,999,487	17,280,065	7,268,305
<i>cite</i>	75,657	421,874	1,604,856	7,712,798	2,400,489

### Cleansing

As Lipczak et al. [2009] pointed out, tags from automatically imported posts are problematic for training and evaluating tag recommenders, since their provenance is unknown. They might have been automatically extracted from the title of a resource or resemble the folder structure of a browser’s bookmark directory and thus do not necessarily reflect the user’s view on the resource. The (in)ability of a recommender to predict such tags does not allow us to draw any conclusion about its performance on predicting user-generated tags. Moreover, in most systems, recommendations are usually not provided during import. Hence, we applied a similar cleansing strategy as described in [Lipczak et al., 2009]: we removed sets of posts that were posted at exactly the same time by the same user. Furthermore for the *cite* dataset, additional cleansing was necessary. A thorough inspection of the data had revealed that the tags *no-tag*, *todo\_mendeley* and (many different) tags like *bibtex-import*, *\*file-import-10-07-11*, or *imported-jabref-library* were frequently (and exclusively) used to indicate imported posts. However, the posts of such an import had not identical but slightly different (consecutive) timestamps and were thus not removed by the above described strategy. Therefore, we additionally removed all posts from *cite* that were exclusively tagged with the tags *no-tag* or *todo\_mendeley*, or a tag matching the regular expression  $\backslash bimport\backslash b$  or  $\backslash bimported\backslash b$  (where  $\backslash b$  indicates a word boundary). In addition, we cleaned all tags as described in [Jäschke et al., 2012]: We ignored tag assignments with the tags *imported*, *public*, *system:imported*, *nn*, *system:unfiled*; converted all tags to lower case; and removed characters which were neither numbers nor letters.

### Properties

The core construction process rapidly reduces the number of tags, users, and resources. For example, from 2,999,487 resources (397,028 tags) in the raw *deli* dataset (cf. Table 7.3) to 588,816 resources (65,050 tags) in the tas-graph-core at level 5. The decline of the number of users for an increasing core level can exemplarily be seen in Figure 7.2(c). The smaller datasets *book* and *publ* quickly vanish with rising core level and although the number of users for *cite* and *deli* is very similar, the number drops much quicker in *cite* than in *deli*. Due to the decrease of the number of nodes, experiments using cores with higher levels require a much lower computational effort



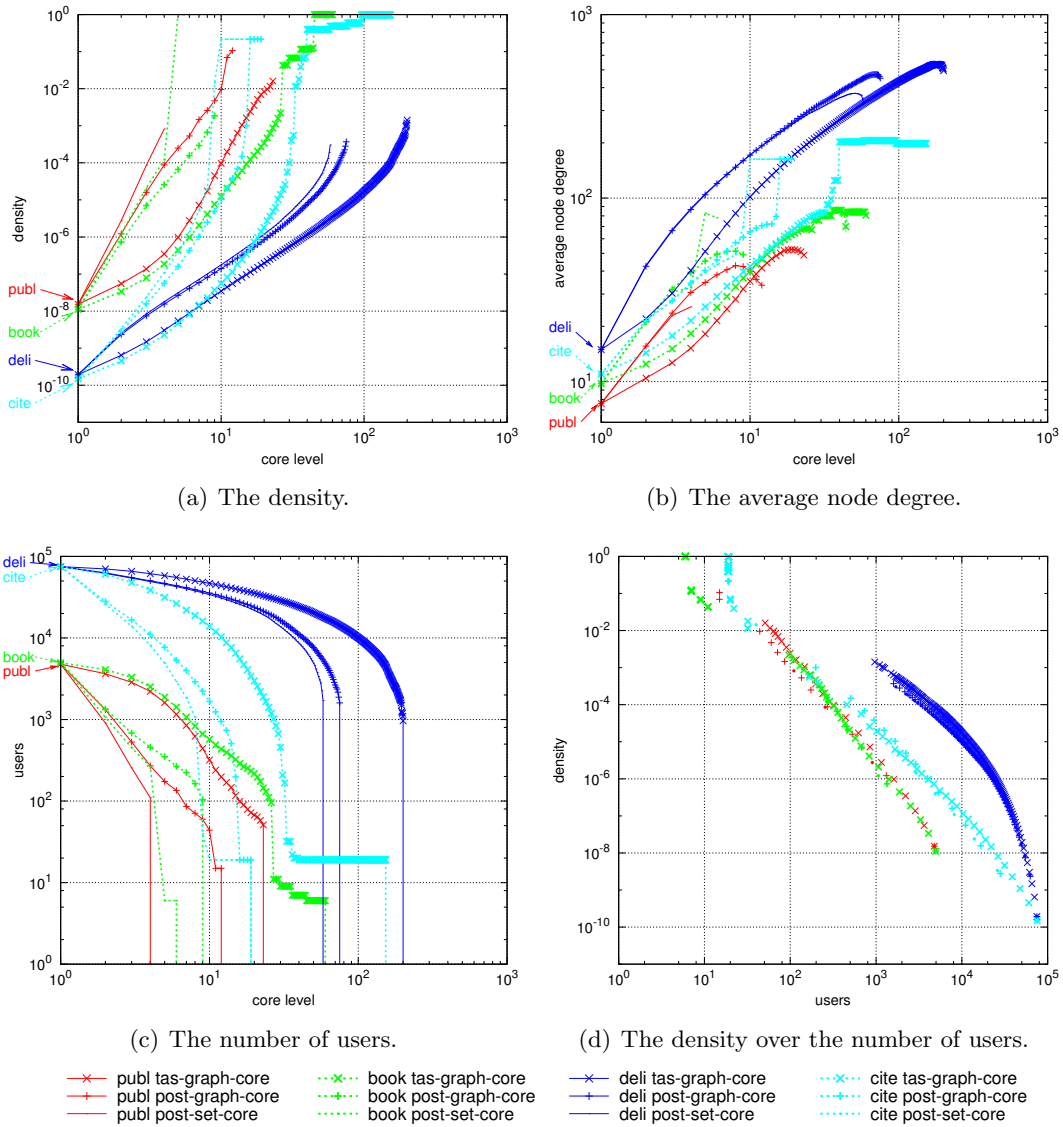


Figure 7.2: The density 7.2(a), the average node degree 7.2(b), and the number of users 7.2(c) in the graphs of the different cores for each dataset as a function of the core level; and the density as a function of the number of users 7.2(d) (all log-log plots).

since the complexity of most recommender algorithms depends on the number of entities (users, tags, and resources) or tag assignments.

Since the usual argument for the use of cores is their higher density compared to raw datasets [Krestel et al., 2009, Ramezani, 2011, Nanopoulos et al., 2013, Seitlinger et al., 2013], we compare this property for all three core types on the four datasets. The *density of a graph* is often defined as the fraction of realized edges among the theoretically possible edges; in [Diestel, 2005] this quantity is also referred to as *edge density*. In a folksonomy, the edge density is equal to  $\frac{|Y|}{|U| \cdot |T| \cdot |R|}$ . Other sources (e.g., Janson et al. [2000]) define the density rather as the ratio between edges and nodes. In that case the density is proportional to the average node degree in the graph. In a folksonomy the average node degree is three times the edge-node ratio:  $\frac{3 \cdot |Y|}{|U| + |T| + |R|}$ , since every hyperedge in  $Y$  connects three nodes. The edge density is also related to the node degree: it can be understood as the ratio between the actual sum of degrees (in a folksonomy that is  $3 \cdot |Y|$ ) and the theoretically possible sum of degrees (in a folksonomy that is  $3 \cdot |U| \cdot |T| \cdot |R|$ ).

Figures 7.2(a) and 7.2(b) show for each dataset and each core type the edge density and the average node degree, respectively, depending on the chosen core level. As expected, the edge density increases with the core level, and for the same level and the same dataset, the post-set-core is the densest core, followed by the post-graph-core and the tas-graph-core. The average node degree at first also rises with the core level, however, with the last levels before the core vanishes, we can observe a decrease for most of the cores. In the case of the tas-graph-core on the *book* dataset, the average node degree drops quickly at core level 44 and then rises again at the next level. This behavior coincides with a sharp drop of the number of users in Figure 7.2(c) and a sharp rise in density in Figure 7.2(a). An inspection of the graph properties showed that from level 43 to 44 the *book* dataset lost almost half of its remaining edges (tag assignments) but only few nodes, while from 44 to 45 it lost two thirds of the edges but also about 75 % of the nodes, resulting in a much smaller and denser graph.

Comparing the behavior with respect to both density and node degree between the three core types, we see that post-graph-cores and post-set-cores are more similar to each other (especially the average node degrees are close together) than to the tas-graph-cores, which always have a lower density and a lower average node degree than the other two types (compared on the same dataset and level). It is also worth noting that the smaller datasets (*publ* and *book*) have fewer users and lower average degrees than the larger datasets (*deli* and *cite*), yet higher densities. This was to be expected: it is a consequence of the number of possible edges (that enters the formula for the density), which rises super-linear with the number of nodes. Among the four datasets and three core types, we can see three cores that reach the maximal density of 1: the post-set-core at level 6 of *book* (at level 5 the density is  $\frac{826}{840} = 0.98\bar{3}$ ), the tas-graph-core at level 45 of *book*, and the tas-graph-core at level 96 of *cite*. There is no general pattern that indicates which main core has the highest density. For

Table 7.4: The levels  $l_m$  of the main cores of the four datasets for the three different core types, and the levels chosen for the experiments. (\* Some levels of the post-set-core were ignored in the evaluation due to the small size of the respective cores, see Section 7.4.2.)

dataset	$l_m$ tas-graph-core	$l_m$ post-graph-core	$l_m$ post-set-core	chosen $l$
<i>publ</i>	23	12	4	2–6
<i>book</i>	60	9	6	2–6*
<i>deli</i>	200	75	58	2–10, 20
<i>cite</i>	153	19	19	2–10, 15*

example, for the *publ* dataset the densest main core is a post-graph-core but for the *deli* dataset it is a tas-graph-core.

Another observation is that, although the density of the raw *cite* dataset ( $0.15 \cdot 10^{-9}$ ) is slightly smaller than that of *deli* ( $0.19 \cdot 10^{-9}$ ), it is growing much quicker with the core level than on *deli* and is already higher than on *deli* at a level of 2 for both the post-set-core and the post-graph-core. Such a rapidly increasing density can be explained by a larger share of sparsely connected nodes compared to well-connected nodes: nodes that are not well-connected are removed in cores of higher levels, while well-connected nodes are more likely to remain in the (thus denser) core. In *deli*, on the contrary, we can infer that the share of well-connected nodes is higher than on *cite*, since the density increases less rapidly. Finally, Figure 7.2(d) confirms that the density increases with a decreasing number of users (and thus with increasing level). We can observe that per dataset the three curves (one for each core type) are almost indiscernible, which indicates that the core type has no pronounced influence on the relation between density and the size of the dataset. Comparing the curves of different datasets, we also note that their slope is a dataset dependent property.

## 7.4.2 Evaluation Methodology

In our benchmarking setup for evaluating different recommenders, we used the *Leave-PostOut* scenario as described here in Section 2.4.2. It is a very common choice in tag recommender evaluation (e.g., Ramezani [2011], Seitlinger et al. [2013], Montañés et al. [2011], Kubatz et al. [2011]). To ensure statistical validity, we repeated each experiment five times – such that every time a post is randomly drawn for each user – and report the averages of the resulting scores.

The dimensions of our experiments are the four datasets, the three different core types, the chosen levels, the recommendation algorithms, and the evaluation metrics.

## Cores

For the experiments we used – besides the raw datasets (or ‘cores at level 1’) – all three types of cores we described in Section 7.2.3. Although the post-set-core allows us to select different thresholds for users, tags, and resources, we used only one single threshold  $l$  for three reasons: (i) to be comparable to the tas-graph-cores and post-graph-cores which do not allow separate thresholds, (ii) to be consistent with most of the previous tag recommender evaluation works without particular focus on special use cases like the consolidation of the tag vocabulary (cf. Section 7.2.3), and, finally (iii) to keep the dimensionality of the experiments manageable.

For each dataset we chose several core levels on which we conducted the experiments (see ‘chosen  $l$ ’ in Table 7.4). The difference in choice is due to the different characteristics of the datasets (size, level of the main core, unchanged cores over several levels, etc.). For the two smaller datasets (*book* and *publ*), we selected five levels (2–6). The two larger datasets (*deli* and *cite*) allow the selection of higher levels and thus we chose consecutive levels up to level 10 and then for each dataset one larger level (20 for *deli* and 15 for *cite*), taking into account that the cores of *cite* vanish much faster than those of *deli*. Due to the rapid rise in density with rising core level, some cores have only very few nodes (cf. Figure 7.2(c)). In particular, the post-set-cores at levels 9 or higher of the *cite* dataset and at levels 5 and 6 of the *book* dataset contain less than 40 users. Such cores do not allow a representative evaluation of recommender algorithms since it would rely on the judgment of very few users. They have therefore been excluded from the analyses. All other considered cores have more than 100 users.

## Evaluation Metrics

The evaluation metric determines the quality of a recommender by measuring how successful an algorithm can predict the tags of the left-out post. We use the two common metrics *recall* and *precision* at a given cut-off level  $k$  ( $\text{rec}@k$  and  $\text{pre}@k$ ). In the experiments we let  $k$  run from 1 through 10. The *mean average precision* (MAP) computes the arithmetic mean of the precision taken at each position of a ranking where the recall changes (cf. Section 2.4.2).

## Recommender Algorithms

Since the goal of our experiments is not to find the best algorithm, but rather to analyze the experimental setup itself, we select a set of well-studied tag recommendation algorithms, namely *most popular tags*, *most popular tags by resource*, *most popular tags by user*, *adapted PageRank*, and *FolkRank* (cf. Section 2.4.3). The two latter algorithms are parametrized and we use the same parameter setting as Jäschke et al. [2008]:  $d = 0.7$  for both the *adapted PageRank* and *FolkRank*. All chosen algorithms are graph-based (in contrast to content-based methods) and thus their performance may depend on the way the folksonomy graph is restricted through the core construction.

Furthermore, we employ the (bogus) *least popular tags* recommender to demonstrate an anomaly that affects the *LeavePostOut* methodology on cores (cf. Section 7.5.4). The algorithm is deliberately designed to produce bogus recommendations by always recommending those tags that occur the least often in the training dataset.

## 7.5 Results

We conducted various experiments to address Research Questions RQ2 and RQ3: In Sections 7.5.1 through 7.5.3 we investigate the influence of the core setup on the performance of recommender algorithms, thus attending to RQ2. In Sections 7.5.4 and 7.5.5, we explain pitfalls of the *LeavePostOut* setting, answering RQ3. In total we conducted 937 experiments using different recommendation algorithms in different setups – each time conducting *LeavePostOut* once for each user. Each single experiment was repeated 5 times and evaluated using 21 different metrics. In the following, we present and discuss our findings:

- We start by summarizing some general results on the performance of recommenders on different cores.
- In Section 7.5.1, we find that the performance of a recommender varies not only over different datasets, core types, and core levels, but also changes when using the same training data and choosing only the test posts from within denser cores.
- Section 7.5.2 addresses the problem of diminished posts, showing that such posts occur frequently in cores and influence the overall performance of recommenders.
- Section 7.5.3 is dedicated to the correlation between rankings of algorithms on different setups. We find that despite high consistency among those rankings, different setups may well lead to different conclusions about the performance of algorithms.
- We point to a statistical flaw of the use of cores within a *LeavePostOut* setup in Section 7.5.4.
- Finally in Section 7.5.5, we demonstrate how the *most popular tags* baseline can be affected by irregular tag distributions.

When we compare recommenders' scores on different cores and levels with different metrics, we first observe that they tend to yield better scores on the post-set-cores than on the post-graph-cores of the same level (in 97.1% of the experiments) and better scores on the post-graph-cores than on the tas-graph-cores (88.0%). The performance of the algorithms on the tas-graph-cores, post-graph-cores, and post-set-cores is better than that on the raw datasets in 94.2%, 99.6%, and 99.7% of the cases, respectively.

This increase of the scores raises the question whether the choice of the core has an influence on the comparison of different algorithms against each other. As an

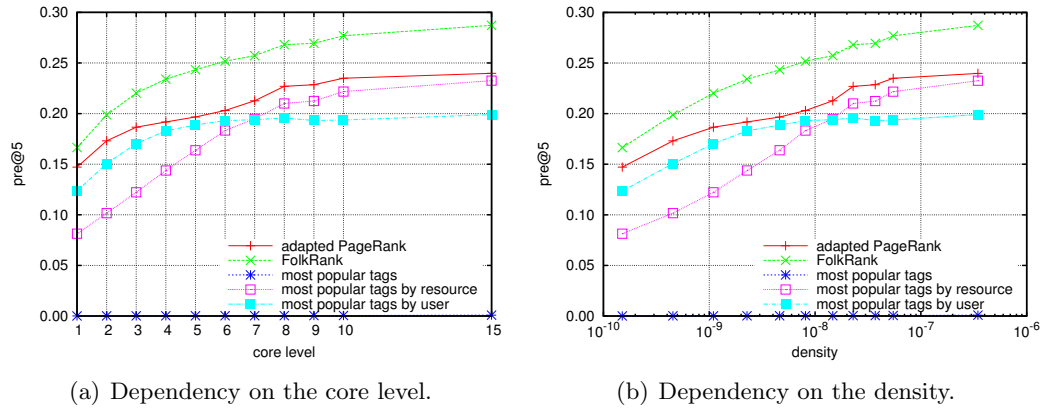


Figure 7.3: An example for benchmarking on different cores: The pre@5 scores of all algorithms on the *cite* tas-graph-core over different core levels and densities.

example how the ranking of algorithms can change with different core levels, Figure 7.3 shows a comparison of the pre@5 of the five algorithms on the *cite* tas-graph-core. The observation that *FolkRank* shows the best performance is in line with prior results [Jäschke et al., 2008]. Although the ranking of the algorithms' performance is quite stable over the levels, the results of *most popular tags by user* are better than those of *most popular tags by resource* on tas-graph-cores for core levels 1 through 6 and worse for higher levels. Thus, a single experiment using the raw data would have yielded another conclusion on the performance of these two recommenders than an experiment using only a core at level 10. We further investigate correlations between such algorithm rankings on different cores in Section 7.5.3.

Figure 7.3 also shows the unexpectedly bad performance of the *most popular tags* recommender on *cite* – a phenomenon we investigate in Section 7.5.5.

### 7.5.1 Recommendation Performance Depends on Core Type and Level

In our experiments, the most prominent observation is that the performance at different core levels depends both on the dataset and on the algorithm – as expected from previous work in recommender systems literature: For example, Cremonesi et al. [2010] found that different algorithms for item (movie) recommendation react differently to manipulation of the test set, while Jäschke et al. [2008] showed for different tag recommenders that their scores vary over datasets of different tagging systems. In Figure 7.4 (the lines labeled “(a) tas-graph-core”), we see exemplarily the pre@5 scores<sup>10</sup> for the five algorithms over different levels of the tas-graph-core for all four datasets. A strong visible tendency is that scores rise with an increasing level –

<sup>10</sup>To suggest five tags is a typical choice in tagging systems. The resulting diagrams for rec@5 and MAP are similar to those for pre@5. They can be found in Appendix C, Figures C.1 and C.2, respectively.

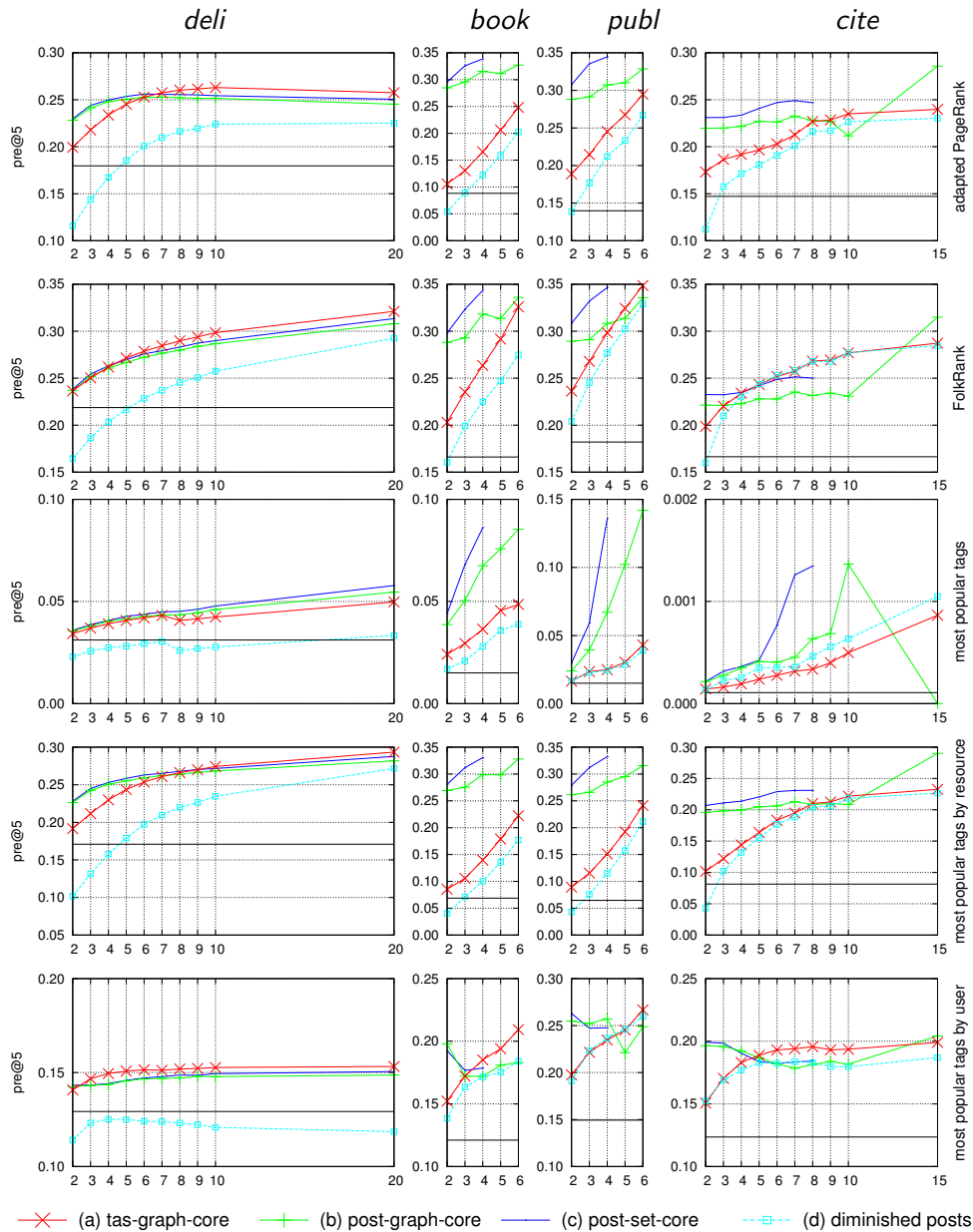


Figure 7.4: The pre@5 scores (on the  $y$ -axis) over the core level  $l$  (on the  $x$ -axis) for *deli*, *book*, *publ*, and *cite* for the five recommenders using modifications of *LeavePostOut*. Each column of plots represents the dataset specified at the top, each row contains results for the algorithm specified at the right, respectively. The horizontal lines depict the pre@5 value for the respective raw dataset.

Table 7.5: Statistics on diminished post in each dataset for several core levels of tas-graph-cores and post-graph-cores. Listed are a core’s share of posts that have lost tags (compared to the original dataset) as well as the average number of tags, that such posts lose. The table shows these statistics for the levels  $l=2, 3$ , and  $l^\top$ , where the latter denotes the highest level that was considered in our experiments:  $l^\top = 6$  for *publ* and *book* and  $l^\top = 20$  and  $15$  for *deli* and *cite*, respectively (cf. Table 7.4).

dataset	share of diminished posts in %						average number of lost tags					
	tas-graph-core			post-graph-core			tas-graph-core			post-graph-core		
	2	3	$l^\top$	2	3	$l^\top$	2	3	$l^\top$	2	3	$l^\top$
<i>publ</i>	18	26	42	17	27	49	1.66	1.82	2.26	1.51	1.60	1.87
<i>book</i>	11	17	31	12	22	40	1.38	1.52	1.85	1.35	1.53	1.77
<i>deli</i>	2	3	7	2	2	7	1.20	1.23	1.29	1.15	1.16	1.22
<i>cite</i>	5	9	30	4	9	41	1.45	1.55	2.30	1.29	1.39	1.75

exceptions are a few levels on *cite* for *most popular tags by user* or *deli* for *most popular tags*.

Further, we leverage the property that the tas-graph-core always contains the post-graph-core, which in turn contains the post-set-core at the same level (Lemma 7.5): Next to the scores on the tas-graph-cores (a) we plotted the scores of the same experiments with only a slight modification of *LeavePostOut*’s post selection process: Where we usually choose one post per user at random, we now choose one post per user randomly such that it is also contained in the post-graph-core (b) or also contained in the post-set-core (c) – scores on diminished posts (d) will be relevant in the next section. Note, that only the selection of the left-out posts is different to (a), as all four variations use the same core (the tas-graph-core) for training. Comparing the scores on arbitrarily chosen posts to those particularly chosen from one of the smaller cores, we see that for most of the algorithms it is easier to predict tags for posts from the post-graph-core than for arbitrarily chosen posts. We yield even better results for posts contained in the post-set-core. The exceptions to that tendency are the same we have observed before. We can conclude that focusing on posts from the dense part of the data often overestimates the performance of recommendation algorithms.

## 7.5.2 Diminished Posts

As already mentioned in Section 7.2.3, diminished posts (i.e., posts having fewer tags in cores than in the raw dataset) are a result of the design of the tas-graph-core and the post-graph-core. In contrast, post-set-cores do not suffer from this issue. To illustrate the influence of such diminished posts, we once more modified *LeavePostOut*’s post selection process (like in the previous section) to randomly choose only such posts (line (d) in Figure 7.4). We can observe that in most cases (with the exception of *most*



*popular tags* on *cite*, a case that is discussed later in Section 7.5.5), the recommenders perform comparably well or worse on posts that have lost tags than on arbitrary posts. Regarding the exception, we have to consider that in general the scores of *most popular tags* are extremely low and thus only very few correctly predicted tags more or less can yield relatively large changes in the scores. The largest difference between the pre@5 scores on arbitrary posts and on diminished posts can be observed on *deli*, for example 0.192 on the full tas-graph-core versus only 0.102 for the diminished posts, with *most popular tags by resource* at level 2. In general, the amounts by which the scores differ are diverse without a clear tendency.

Table 7.5 shows that diminished posts are not only a theoretical problem, but do indeed occur frequently in cores. We can see that on the two smaller datasets (*publ* and *book*) even for level 2 more than 10% of the posts in the core have lost tags, while there are fewer such posts in the larger datasets. Raising the core level, however, raises the share of diminished posts – most dramatically in the *cite* dataset, that has 5% diminished posts in the tas-graph-core (4% in the post-graph-core) at level 2 but a share of 30% (41%) at level 15, the highest level used in our experiments. The *deli* dataset has significantly lower shares of diminished posts, yet also shows the tendency of a rising share for a rising core level. The second half of Table 7.5 shows the average number of tags that diminished posts have lost. Again, we can observe that the numbers rise with a rising level. Each such post loses one or two (and even more in the higher levels of the tas-graph-core) tags on average. Lost tags pose an artificially introduced difficulty to the evaluation of tag recommendation algorithms, as there are less correct tags that could be predicted. Especially with a metric like pre@5, one or two more tags to predict can make an enormous difference.

These observations support the assumption that diminishing posts has indeed an influence on the evaluation and is thus undesirable, as it is not clear how different algorithms react to such artificially modified posts. A reason for the weaker performance might be that it is easier to yield a higher precision when there are more tags to predict and thus it is more likely that one of these tags is recommended. However, we could observe the same behavior for the rec@5 and MAP scores (with even fewer exceptions).

### 7.5.3 Recommender Ranking Correlation

The goal of evaluating recommender systems usually is to determine one algorithm that performs best on one or more datasets, and therefore several algorithms are ranked according to their performance. Since various setups for experiments are possible – several core types, levels, and metrics – the question arises, whether the ranking of recommenders varies depending on the chosen setup. To investigate this question we determine the algorithm rankings where the algorithms are ranked according to their recommendation quality. A ranking can be computed on the raw datasets, on all

Table 7.6: Inconsistent rankings of recommenders over various benchmarking setups: Shown are the mean pairwise Pearson’s  $r$  and the number of discordant pairs  $d$  in the recommender algorithm rankings on different cores together with their standard deviation  $\sigma$ .

dataset/metric	avg. $r$	$\sigma$	avg. $d$	$\sigma$
<i>publ</i>				
MAP	0.912	0.074	1.473	1.148
pre@5	0.909	0.079	1.593	0.977
rec@5	0.920	0.076	1.516	1.026
<i>book</i>				
MAP	0.908	0.092	1.429	1.087
pre@5	0.878	0.117	1.462	1.148
rec@5	0.912	0.090	1.330	1.076
<i>deli</i>				
MAP	0.994	0.008	0.512	0.500
pre@5	0.992	0.010	0.361	0.481
rec@5	0.993	0.009	0.503	0.501
<i>cite</i>				
MAP	0.981	0.026	0.651	0.735
pre@5	0.972	0.043	0.492	0.676
rec@5	0.976	0.043	0.595	0.766

three core types, and at all chosen levels.<sup>11</sup> Between two rankings (on two different setups), we can determine Pearson’s correlation coefficient  $r$  (see Section 2.1.1), as a measure of how likely the score rankings of the recommenders are (linearly) correlated (cf. Section 2.1.1). As Pearson’s  $r$  takes the particular score values (the value describing one recommender’s performance on one setup) of the algorithms into account, we additionally use another metric that only considers the order of the algorithms in a ranking: the *number of discordant pairs*  $d$ .<sup>12</sup> Given two rankings, the algorithms  $A$  and  $B$  are *discordant*, when in one ranking  $A$  performs better than  $B$  while in the other ranking  $B$  is better than  $A$ . Thus in our case of five algorithms,  $d$  is between 0 (the rankings agree completely) and 10 (one ranking is the reverse of the other).

Table 7.6 shows the mean pairwise (averaged over any pair of two different setups) values of  $r$  and  $d$  together with the standard deviations exemplarily for the metrics

<sup>11</sup>That is, 14 different setups for *book* and *publ* each and 28 and 31 setups for *cite* and *deli* respectively.

These numbers are determined by the choice of levels (see Table 7.4) and the exclusion of cores with only few users (see Section 7.4.1).

<sup>12</sup>The number of discordant pairs is closely related to the ranking correlation measure *Kendall’s*  $\tau$ . In fact, since all rankings have the same length of five (algorithms), and no two algorithms have equal scores in one ranking, we have  $\tau = 1 - 0.2d$ .

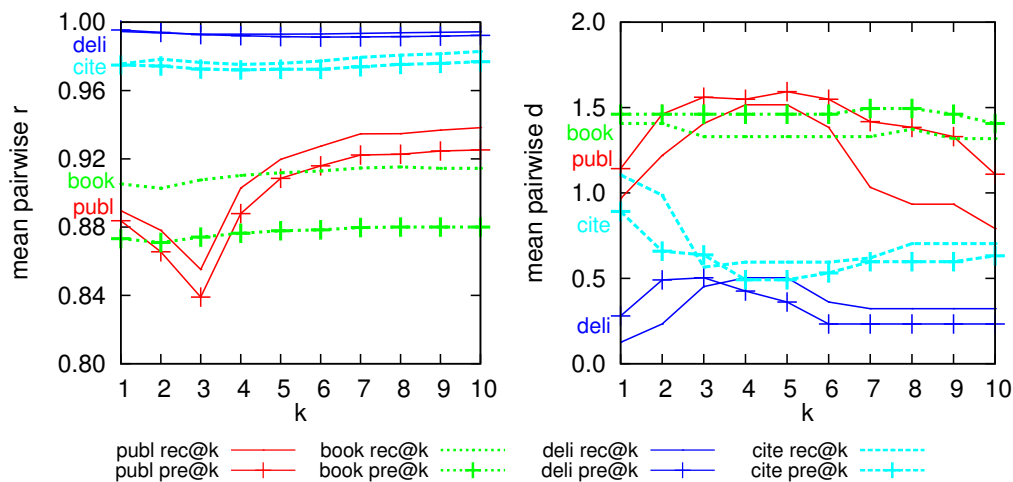


Figure 7.5: Inconsistent rankings of recommenders over different cut-levels  $k$  of precision and recall: Shown are the mean pairwise Pearson’s correlation  $r$  and number of discordant pairs  $d$  over the cut-level  $k$  for the metrics  $\text{rec}@k$  and  $\text{pre}@k$ .

$\text{pre}@5$ ,  $\text{rec}@5$ , and MAP. We can observe that on no dataset, we get perfect correlations. Generally, the correlations are rather high, but we clearly see that the rankings are inconsistent. The most stable are the rankings on *deli*. Here, only in every second pair of setups, two recommenders change their order. The correlations on *cite* are only a little lower than on *deli* and on the two BibSonomy datasets the values are similar and again lower than those on *cite*: on average, in two rankings one or two pairs of recommenders have a different order.

Further, we computed which of the cores yield the ranking that is most consistent with the raw data. For all datasets these are the *tas-graph-cores* at levels 2 and 3 (i.e., the two largest cores). More generally, we could observe that higher levels (and thus higher densities) tend to yield results less consistent with the raw data. We conclude that in experiments, cores with lower levels are preferable to others, since they resemble the original dataset more closely.

The consistency of the rankings also depends on the particular metric that is employed. In Figure 7.5, we see the mean pairwise values of  $r$  and  $d$  for  $\text{rec}@k$  and  $\text{pre}@k$  with  $k$  running from 1 through 10. Clearly, the behavior of the consistency measures over the levels is dataset-specific. For *deli*, the consistency is quite stable for both metrics, precision and recall. However, for the two BibSonomy datasets (most notably *publ*), the values vary, and the highest consistency is achieved for  $k = 10$ , indicating that especially among top recommendations the recommenders’ success changes with the setups. Finally on *cite*, the difference in consistency is more dramatic when measured by the number of discordant pairs than with  $r$ . This means that recommenders switch places in the performance rankings although their scores develop similarly with changing levels or core types.

Table 7.7: For each dataset and each core type, the core levels  $l$  (among those considered in the previous experiments, see Table 7.4) where the algorithm *most popular tags* outperforms *least popular tags* (column "mpt") or otherwise (column "lpt"), according to precision at five (pre@5) and recall at five (rec@5). For both metrics, the comparisons are identical except for the tas-graph-core of *publ* at level 5. The difference is indicated using the metric as superscript.

dataset	tas-graph-core		post-graph-core		post-set-core	
	mpt	lpt	mpt	lpt	mpt	lpt
<i>publ</i>	1, 5 <sup>rec</sup> , 6	2 – 4, 5 <sup>pre</sup>	1, 3 – 6	2	1, 3, 4	2
<i>book</i>	1, 4 – 6	2, 3	1, 3 – 6	2	1 – 4	–
<i>deli</i>	1 – 10, 20	–	1 – 10, 20	–	1 – 10, 20	–
<i>cite</i>	1	2 – 10, 15	1, 4 – 10, 15	2, 3	1, 4 – 8	2, 3

#### 7.5.4 Exploiting Cores Using LeavePostOut

To demonstrate a critical statistical flaw of the use of any core in connection with the *LeavePostOut* method, we employ the bogus *least popular tags* recommender, that always suggests the rarest tags. It is expected that this method's scores should be inferior to those of the other algorithms. They are indeed, when the raw datasets are used – recall and precision always yield 0 and the MAP-score is below  $10^{-4}$ . However, on the two BibSonomy datasets and on *cite*, this changes once we use cores at a level  $l > 1$ : On many of the investigated cores, *least popular tags* actually outperforms *most popular tags* (and occasionally even *most popular tags by resource*). Table 7.7 shows for the three core types those levels on which *least popular tags* yields better scores – measured in terms of precision and recall – than *most popular tags* or the other way around.

The algorithm *least popular tags* can profit in cases where the left-out post contains many rare tags: through *LeavePostOut*, these tags become even rarer and, in particular when they occurred exactly  $l$  times in a core at level  $l$  before *LeavePostOut*, they occur  $l - 1$  times afterwards. Instead of being removed from the dataset like in the case of  $l = 1$  (the raw data), for higher levels  $l$ , they become the rarest tags in the core. We can observe that this effect is mitigated with an increasing core level and the scores of *least popular tags* tend to fall below those of *most popular tags*. Only on *deli* *least popular tags* is always worse than *most popular tags* (although its scores are still significantly higher than zero). This can be explained by a much higher average number of tag assignments per tag: 43.5 on *deli* compared to only 6.9, 12.8, and 18.3 on *publ*, *book*, and *cite*, respectively. The higher the number, the less likely it is to select posts with tags that occur exactly  $l$  times during *LeavePostOut*. The same argument explains why *least popular tags* falls behind *most popular tags* as the level increases: together with  $l$  also the average number of tag assignments per tag rises.

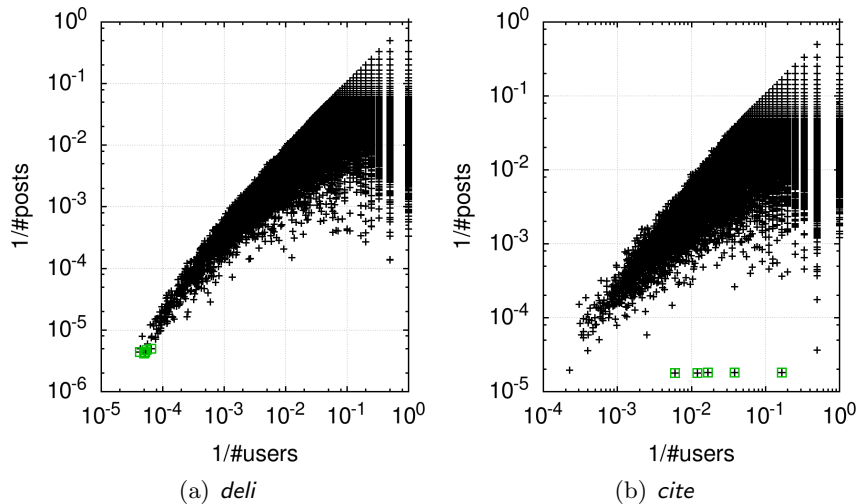


Figure 7.6: The distribution of the number of users per tag versus the number of posts per tag. Shown are the reciprocal values in a logarithmic plot to focus on the high-frequency tags. The top five tags of the *most popular tags* recommender are highlighted.

### 7.5.5 The Most Popular Baseline

In Section 7.5.1, we have seen that the *most popular tags* recommender performs very badly – especially on *cite*. To explain this phenomenon, recall that the most popular tags are computed based on their *post frequency* (i.e., the number of posts that contain a tag). Thus, if there are tags that are used extremely often by only a few users, they will be among the most popular tags and therefore be recommended to many users. In Figure 7.6, for *deli* and *cite*, for each tag, its post frequency is plotted against the number of users that have tagged at least one post with it. To put emphasis on those tags that occur most often, we have plotted the reciprocal values – and thus small values correspond to high frequencies – on a log-log scale. Relevant for the *most popular tags* recommender are the tags with the highest post frequency – these are the ones closest to the  $x$ -axis. We can see that the top five tags for *deli* are also close to the  $y$ -axis, which means they have both a high post frequency and a high user frequency. In contrast, the tags with the highest post frequency in *cite* have a rather low user frequency; therefore they are a bad recommendation for most of the users. A closer look at these top five tags (namely “celegans”, “elegans”, “nematode”, “caenorhabditiselegans”, and “wormbase”) reveals that they are all related to *Caenorhabditis elegans*,<sup>13</sup> a worm which is frequently used as a model organism in biology. These five tags were very frequently used (27,735 times) in the same posts by two users (with IDs 33569 and 28123) and less frequently by 165, 81, 58, 24, and

<sup>13</sup>[http://en.wikipedia.org/wiki/Caenorhabditis\\_elegans](http://en.wikipedia.org/wiki/Caenorhabditis_elegans)

4 other users, respectively. Thus, the posts from the two users were likely created automatically but were not detected by the approaches described in Section 7.4.1, since they describe the content of posts and not their creation process (as do tags like “imported”). Also, if we recall the evaluation procedure *LeavePostOut*, which randomly picks one post for every user, it becomes clear that these most popular tags are only a weak baseline. A better choice would be to measure the popularity of tags based on the number of users who used it at least once, instead of on the number of overall tag assignments.

## 7.6 Conclusion

We have analyzed the use of cores for the evaluation of tag recommendations. The main contribution of this chapter is the extension of the framework of core constructions through the introduction of set-cores as generalization of graph-cores. This new core type allows us to transfer the idea of cores to datasets without imposing a graph structure on them. In contrast to graph-cores, it allows the use of several thresholds at once and for flexible combinations of individual and combined thresholds for different entities. We have successfully used them in tag recommender benchmarking experiments to avoid the problem of diminished posts, answering Research Question RQ1. Regarding RQ2, we have shown that using cores in recommender benchmarks has an influence on the results and regarding RQ3, we saw two pitfalls that arise from the use of *LeavePostOut* to select training and test data.

### 7.6.1 Lessons Learned

In the experiments, we have confirmed that benchmarking results do not only depend on the dataset and preprocessing procedures but also on the chosen cores, and that using cores for offline evaluation has its pitfalls: The use of tas-graph-cores and post-graph-cores results in diminished posts (post with fewer tags than they had originally) in the dataset. With the use of post-set-cores, we have presented a suitable solution for this problem. The anomaly of the successful *least popular tags* recommender directly exploits the combination of cores and *LeavePostOut*. For other recommenders, it is unclear whether and how they can profit from the particular setup or the artificial rareness of the left-out tags. We have also confirmed that recommenders perform differently in different core setups of the same dataset. Thus, focusing on one particular core can produce non-stable results. Evaluating the performance of recommenders on another core type or at another core level might cause changes in the results. There is no guarantee that a recommender performing best in one setup is also the best in another setup (even on the same dataset). The correlations of recommender rankings over various setups were relatively high. Yet, the fact that in a comparison of different algorithms, some of them switch ranks on different cores suggests that the choice of the core and its level is even more critical for the comparison of algorithms with similar

performance and for the optimization of parametrized algorithms (where usually scores change only little through fine adjustments of parameter values).

### 7.6.2 Recommendations for Future Tag Recommender Benchmarking Experiments

Following our findings, we can draw the following conclusions for future experiments with recommender algorithms:

- In general, the comparison of tag recommender algorithms should always be performed directly on the raw data or on several core types and levels.
- Differences in the rankings, resulting from such comparisons, indicate strengths or weaknesses of individual algorithms in the presence of datasets with different densities.
- We could observe that even cores at higher levels still yield correlated results to those of the raw data. It is therefore worth comparing recommenders on several of these smaller subsets of the raw data to get a first impression of their overall performance, before running the computationally more expensive experiments on the raw data.
- We suggest to still use small choices for the core level (thus larger cores), since they yield more consistent results with the raw dataset.
- We recommend not to run an evaluation on only one arbitrary chosen core, but to carefully select several levels that suit the investigated use case. The particular choice of the core level should be motivated by the use case – examples are given in Section 7.2.3.
- To avoid the problem of diminished posts, post-set-cores should be used. Investigating posts with all their tags (as they are in the raw data) is closer to the real online use case. Allowing diminished posts increases the divide between offline evaluation and actual online usage further.
- To tailor the test dataset to a particular use case, post-set-cores – in contrast to tag-graph-cores and post-graph-cores – allow to impose thresholds individually for each dimension of the data.

### 7.6.3 Future Research

We have shown that the choice of core type and core level has an impact on a benchmarking experiment's result. It is well-known that there are many other parameters of the experimental setup which are influential as well. While, for instance, the choice of the evaluation metric can often be justified by the use case – for example, by the design of the service in which the recommendations are provided – other choices are

often rather arbitrary or for the sake of minimizing the computational effort. These aspects of the experimental setup include the method of splitting test and training data (e.g., *LeavePostOut*, different variations where more than one post is removed from the dataset, or time splits), the sampling of the training data (e.g., selecting some randomly chosen post per user, or selecting the most recent post of each user, or using some user-independent selection of posts), and the preprocessing of the data. Further experiments could reveal the influence of these choices on the results of tag recommender benchmarking as well as insights about how particular algorithms can profit or suffer from the chosen setup.

The fact that different core setups yield different recommender rankings is an indicator that different algorithms have strengths and weaknesses when dealing with rather sparse or with more dense data. In Section 7.5.1, we have shown that performance differences occur even through choosing only the test posts (the user-resource combinations to recommend tags for) from different regions of the data (i.e., from the *tas-graph-core*, the *post-graph-core*, or the *post-set-core*) while leaving the training dataset the same (i.e., in our experiments the *tas-graph-core*). This encourages approaches using different recommender algorithms in different situations: A recommender that performs well (compared to others) on sparse data can be applied to new (or sparsely connected) users and resources. An algorithm that dominates on the more dense datasets (higher core levels) can be chosen for user-resource pairs from an already dense section of the data. The dynamic selection of the appropriate recommender – depending on the user and resource at hand – can be investigated as a machine learning problem.

An open aspect regarding the offline evaluation of a recommender is the current way of distinguishing good recommendations from bad ones. In the current setup, a tag is only a good suggestion if it fits the user’s actual choice exactly (in our experiments ignoring upper and lower case). Thus, for example, the recommended tag “work” is considered a bad recommendation even if the user had in fact used the seemingly related tag “working”, while in an online setting, the user might have accepted the recommended tag. Several approaches to “softer” measures are conceivable: word stemming of both recommended tags and the actual tags (actually the conversion to lower case is already a mild form of stemming), differentiating between exact and close fits of recommended tags, and so on. In Section 7.3 we have already mentioned the approach by Mishne [2006] of using string distance to compare tags. Different evaluation scenarios could be compared in a similar setup, like in this chapter, varying the evaluation functions instead of core type and level. Such experiments should also be accompanied by a user-study to investigate, for instance, which forms of stemming are acceptable for many users.

Finally, since set-cores can be constructed on arbitrary sets, they can be used in the analysis of all kinds of datasets. In the related work on cores in Section 7.3, we have mentioned several applications of graph-cores for diverse purposes, such as community detection, data visualization, or the discovery of dynamics in datasets. It is now possible to adapt these methods using set-cores, and thus, to extend their scope and their analytic capabilities – through new, flexible, and multi-valued property functions.



## Chapter 8

# ◆ Folksonomic Recommendation of Scholarly Literature



The ever-growing flood of new scientific articles requires retrieval mechanisms that go beyond traditional full-text search. To mitigate this instance of the information overload problem, social bookmarking systems employ recommendation algorithms to present personalized lists of interesting and relevant publications to their users.

In this chapter, we analyze different ways to incorporate social data and metadata from social bookmarking systems into the graph-based ranking algorithm *FolkRank* to utilize it for recommending scholarly articles to users. We critically review *FolkRank*, explain how it can be extended, and then demonstrate the different variants of *FolkRank* on datasets of the scholarly bookmarking system BibSonomy. We compare their results to those of *collaborative filtering*, which has previously been applied for resource recommendation. In this chapter, we present modified<sup>1</sup> and extended versions of the studies published in [Doerfel et al., 2012a] and [Doerfel et al., 2013a].

### 8.1 Introduction

Of particular importance for every researcher are *scientific publications*. However, especially during the last years, the ever faster growing number of published articles (e.g., Price [1963], Bornmann and Mutz [2015]) has led to the well-known phenomenon of *information overload*. It has become harder and more time-consuming for researchers to keep track of the important publications in their respective fields or to assemble comprehensive “related work” sections for a new article. The search for previously published material is often conducted on the web, using specialized search engines, editorially controlled scientific databases, or systems of user-generated content on the matter of interest. To the latter belong scholarly bookmarking systems. However, even in such systems, the number of publications posted by their users makes it more and more difficult to find or stumble upon interesting articles.

One solution for this problem are *recommender systems* that try to suggest interesting and relevant content to the user. In this chapter, we focus on the recommendation of

---

<sup>1</sup>In contrast to the two previous studies, here we use a different version of *FolkRank* that avoids the convergence issues that we discuss in Section 8.2.3. The results are similar.

scientific publications to users of the scholarly social bookmark and publication sharing system BibSonomy (see Section 2.3.2). That is, given a user, we aim to provide a ranked list of publications that might be of relevance to them. Thus, in our discussion of scholarly social bookmarking in this thesis, here, we approach the task of actively supporting researchers by pointing them to relevant publications.

Folksonomic recommendation algorithms usually employ data mining methods to gain information on the content of resources or to leverage the wisdom of the crowds, by mining the folksonomy structure, for personalized suggestions of resources. To the latter kind belongs the particularly versatile folksonomic recommendation algorithm *FolkRank* [Hotho et al., 2006c]. *FolkRank* was found to be a well-performing algorithm for tag recommendation [Jäschke et al., 2008], but also to be able to identify trends [Hotho et al., 2006c], to produce topic-specific rankings [Hotho et al., 2006c], or even to discover hierarchical relations between tags [Cattuto et al., 2008]. Therefore it is as well a favored candidate for recommending resources.

While *FolkRank* runs on the folksonomy structure and is thus independent of the resource type, the question arises, whether additional information can be included into the algorithm to improve its performance. In scholarly tagging systems, like BibSonomy, not only the data from a publication's usage (the folksonomy structure) but also some metadata of a publication is often available. It is entered by users when they post resources, but it is often sparse and does rarely include a publication's full text.

In this chapter, we adapt the graph structure underlying *FolkRank* and we change its mode of personalization to add additional information: We experiment with both ways to augment *FolkRank* with metadata of publications, as well as with further usage data, such as similarity between users or tags, user groups, or the recently posted publications. Different extensions of *FolkRank* are used in an offline benchmarking for resource recommendation algorithms to investigate how the incorporation of additional knowledge about publications and users can improve the recommendation quality of *FolkRank*.

**Research Questions.** By providing helpful recommendations of relevant literature to researchers, algorithms like *FolkRank*, can help researchers identify those publications that are useful for their research. Thus, advancing these algorithms to produce better recommendations is a step forward towards active support of researchers during their literature research. The subjects of investigation in this chapter are the recommendation algorithm *FolkRank* and its possible extensions. Our research questions are:

- (RQ1) How can *FolkRank* be modified to include further data that is available in a scholarly social bookmarking system?
- (RQ2) Which modification of *FolkRank* can boost its quality as a recommender for scholarly publications?

**Contributions.** The main contribution of this chapter is an extensive analysis of various ways to improve the success of the *FolkRank* algorithm for recommending scholarly publications.

1. We recall previous versions of *FolkRank* and resolve some confusion about its parameters.
2. We show and demonstrate two ways of modifying *FolkRank* to include further data, beyond the folksonomy structure, which is its usual input. Particularly, we discuss the inclusion of metadata of the scholarly publications – data from the creation phase of the publication life cycle –, as well as leveraging data from the usage of the tagging system – data from the usage phase of the publication life cycle –, such as similarity between users and recent interests.
3. We compare different versions of *FolkRank* to each other and to three baselines in a resource recommender scenario, using publicly available data from the real-world scholarly bookmarking system BibSonomy.

**Limitations.** Our investigation covers data from the social bookmarking system BibSonomy – among it the same dataset that has been previously used in a large comparison of various folksonomic resource recommender algorithms by Bogers [2009]. We use different set-cores (introduced in Chapter 7) and two datasets from BibSonomy to make sure that our results do not depend on the particular preprocessing. Yet, to yield overall conclusive results on the performance of the *FolkRank* extensions, one would need a comparison over various datasets. However, not all of the proposed types of data are available on other systems. For all experiments, we speculate about possible explanations for their success or failing. Since these explanations are not BibSonomy specific, they might also hold in comparable systems.

**Structure.** This chapter is structured as follows: We start our investigation with a recall of *FolkRank* in Section 8.2. In Section 8.3 follows a review of related work. Then, in Section 8.4, we describe the experimental setup and the datasets underlying our analysis. The results of our study are presented in Section 8.5. We conclude with an outlook on future work in Section 8.6.

A slight variation of this chapter’s study has previously been published as [Doerfel et al., 2013a] and before that as [Doerfel et al., 2012a], where we used a different version of *FolkRank*. Extending the previous material, we have added the discussion on different versions of *FolkRank* in Section 8.2 which also explains the choice of a different version here. The experiments have been extended and the findings have been rearranged for this thesis.

## 8.2 FolkRank

The central algorithm in this chapter is *FolkRank*, a folksonomic recommendation algorithm, that, similar to *PageRank* [Brin and Page, 1998], uses the structure of a graph between the entities of a folksonomy. We begin this section by reviewing the algorithm, before we introduce extensions which allow us to include further data beyond the plain folksonomy structure which is the basis of *FolkRank*.

### 8.2.1 A Recall of FolkRank

The *FolkRank* algorithm has been created by Hotho et al. [2006c], who applied the algorithm to produce personalized recommendations of users, tags, and resources. It was again presented in a slightly simplified version in [Hotho et al., 2006a], where it was used to identify trends for an entity over time, regarding a specific topic. In the following, we recall this simplified version; the original is however mentioned below, when we describe the *adapted PageRank*.

*FolkRank* consists of two steps: an adaptation of the graph structure and a differential approach between a personalized and an unpersonalized *PageRank*.

#### The FolkRank Graph

First, the folksonomy  $\mathbb{F} = (U, T, R, Y)$  (see Section 2.3.1 for the definition of the folksonomy model) is converted into an undirected, weighted, tri-partite graph  $G_{\mathbb{F}} = (V, E)$ , where the node set  $V$  is the disjoint union  $V = U \cup T \cup R$ , and each tag assignment  $(u, t, r) \in Y$  yields three edges in  $E$ :  $\{u, t\}$ ,  $\{u, r\}$ , and  $\{t, r\}$ . The weighting function

$$w : E \rightarrow \mathbb{N} : \{x, y\} \mapsto \begin{cases} |\{(x, y, r) \in Y\}| & \text{if } x \in U \text{ and } y \in T \\ |\{(x, t, y) \in Y\}| & \text{if } x \in U \text{ and } y \in R \\ |\{(u, x, y) \in Y\}| & \text{if } x \in T \text{ and } y \in R \end{cases} \quad (8.1)$$

assigns to each edge the number of tag assignments that it represents according to the rule above.

#### The Adapted PageRank

The *adapted PageRank* is a personalized *PageRank*, computed on the *FolkRank* graph  $G_{\mathbb{F}}$ . In terms of the original *PageRank* setting, each entity in  $V$  is interpreted as a web page and each edge  $\{x, y\}$  in  $E$  as two links: one from  $x$  to  $y$  and one the other way around. The result of the algorithm is a weight vector  $\vec{w}$ , where each entity in  $V$  corresponds to one entry in  $\vec{w}$ . The entry  $\vec{w}_i$  is the score of the entity  $i \in V$ .<sup>2</sup> Thus  $\vec{w}$  contains a ranking of all entities in the folksonomy, which can easily be separated into three rankings, one for each type of entities (users, tags, and resources).

---

<sup>2</sup>Here, we presume some linear order on  $V$ .

The *adapted PageRank* – according to Hotho et al. [2006a]<sup>3</sup> – is now iteratively computed as fix point of the equation

$$\vec{w} \leftarrow dA^T\vec{w} + (1-d)\vec{p}, \quad (8.3)$$

with the following notation:

- $A$  is the row-stochastic version of the adjacency matrix of  $G_{\mathbb{F}}$ , more formally:  $a_{ij} = \frac{w(\{i,j\})}{\text{str}(i)}$  if  $\{i,j\} \in E$  and  $a_{ij} = 0$  otherwise (where  $\text{str}(i)$  is the strength of node  $i$  in  $G_{\mathbb{F}}$ , see Section 2.1.2).
- $\vec{p}$  is a preference vector, which like  $\vec{w}$  has one entry for every entity in  $\mathbb{F}$ . Setting all entries of  $\vec{p}$  to the same value will yield a global ranking without personalization. Choosing higher values for some particular entries in  $\vec{p}$ , corresponding to some *active* entities, will yield a personalized ranking for these entities. It is required that  $\sum_i \vec{p}_i = \sum_i \vec{w}_{0,i}$ , where  $\vec{w}_0$  is the initial setting of  $w$  in 8.3. Thus, the sum of all weights in  $\vec{w}$  will remain constant during the iteration.
- $d \in [0, 1]$  is a parameter to determine the influence of  $\vec{p}$ .

### The FolkRank

In contrast to the graph of *PageRank* in the setting of web pages, the graph  $G_{\mathbb{F}}$ , constructed from a folksonomy  $\mathbb{F}$ , is undirected. In the web setting, a link’s interpretation is that the linking page marks the linked page as relevant. In the folksonomy, this relationship is less pronounced as each link (edge) between two entities works both ways. Thus, two connected entities mark each other as relevant. Hotho et al. [2006c] therefore introduced a differential approach particularly for the undirected setting: After choosing a suitable preference vector  $\vec{p}$  (see below), two rankings are computed as fix points of Equation 8.3:

- The *adapted PageRank*  $\vec{w}^1$  is computed using  $0 < d < 1$ .
- An unpersonalized baseline ranking  $\vec{w}^0$  is computed, setting  $d = 1$ .
- The final weight vector  $\vec{w}$  – the result of *FolkRank* – is the difference of the these rankings:  $\vec{w} = \vec{w}^1 - \vec{w}^0$ .

*FolkRank* thus computes the “winners and losers” [Hotho et al., 2006c] of the personalization compared to an unpersonalized baseline.

<sup>3</sup>Before, *adapted PageRank* was introduced in [Hotho et al., 2006c] as result of the iteration

$$\vec{w} \leftarrow \alpha\vec{w} + \beta A^T\vec{w} + \gamma\vec{p}, \quad (8.2)$$

requiring  $\alpha + \beta + \gamma = 1$  and  $\alpha, \beta, \gamma \in [0, 1]$ . It is easy to see that this iteration can be transformed into 8.3, setting  $\alpha = 0, \beta = d$ , and  $\gamma = (1-d)$ ; or the other way around, setting  $d = \frac{\beta}{(1-\alpha)}$ . For the latter, one must require  $\alpha \neq 1$ , which is, however, not a real restriction as for  $\alpha = 1$  the iteration would degenerate to  $\vec{w} \leftarrow \vec{w}$ .

## 8.2.2 The Preference Vector $\vec{p}$

In the preference vector  $\vec{p}$ , high weights should be assigned to those entities for which the ranking is computed. To generate a ranked list of tags for a single active entity (a tag, a user, or a resource), Hotho et al. [2006c] assigned a weight to the respective entity and zero to all other values. Hotho et al. [2006a] computed *FolkRank* by assigning preference to specific tags: The weights were distributed such that the active tags received 50% of the total weight, and the rest was spread uniformly over all other entities. Jäschke et al. [2008] used *FolkRank* for tag recommendations and thus had to create a personalization for two entities at once: the active user  $u$  and the resource  $r$  to be posted. They set  $\vec{p}_u = 1 + |U|$ ,  $\vec{p}_r = 1 + |R|$ , and  $\vec{p}_i = 1$  for any other entity  $i \neq u, r$  (before normalizing  $\vec{p}$  such that  $\sum_i \vec{p}_i = \sum_i \vec{w}_{0,i}$ ). The most notable difference in the preference selection between [Jäschke et al., 2008] and [Hotho et al., 2006a] on the one hand, and [Hotho et al., 2006c] on the other hand is the value in  $\vec{p}$  for inactive entities, which is either set to some small value (compared to that of the active entities) or to zero.

Jäschke et al. [2008] also modified the setup in another way, which we address in the next section.

## 8.2.3 Convergence and Variation of FolkRank

In the above mentioned scenario, the initial weight vector  $\vec{w}_0$  is not considered a parameter since the iterations should converge to a unique solution independent from  $\vec{w}_0$ . Indeed, for  $0 < d < 1$  convergence is assured. We sketch a proof, following the argumentation of Bianchini et al. [2005], who proved convergence for the regular *PageRank*, where the entries in the preference vector are equal for all entities. To that end, we refer to three results on matrix algebra that can be found in [Golub and Loan, 1996] and that have been repeated for the sake of self-containedness in Appendix D: By construction, the columns of  $A^T$  sum up to one and thus  $\|A^T\|_1 = 1$  and  $\|dA^T\|_1 = d$ . From Lemma 2.3.3 [Golub and Loan, 1996] follows that  $I - dA^T$  is non-singular ( $I$  being the identity matrix of the same dimensions as  $A^T$ ). Further, by Theorem 7.2.1 (ibid.) follows that all eigenvalues of  $dA^T$  are smaller or equal to  $d$ . Thus for  $0 < d < 1$ , Theorem 10.1.1 (ibid.) is applicable and we conclude that the iteration in Equation 8.3 converges for any starting weight vector  $\vec{w}_0$  to the fix point  $\vec{w} = (I - dA^T)^{-1}\vec{p}$ .

This result is however not applicable to the baseline computation  $\vec{w}^0$  where  $d$  is set to  $d = 1$ . Thus for the baseline, convergence might depend on the initial weight vector. This was also observed by Kim and El Saddik [2011], who showed for a toy example that convergence indeed varies with the choice of the initial weight vector. The example presents a folksonomy graph that is not connected. Indeed, connectedness is the criterion for convergence in the case  $d = 1$ : Here, Iteration 8.3 simplifies to  $\vec{w} \leftarrow A^T\vec{w}$ , describing a Markov chain. Such an iteration converges independently of the initial vector if the Markov chain is *aperiodic* and *irreducible* (compare Section 2.4

in [Brandes and Erlebach, 2005]). As long as every entity in the folksonomy is part of at least one tag assignment, the chain is aperiodic.<sup>4</sup> Since usually, isolated nodes in a folksonomy are ignored,<sup>5</sup> the only requirement is the irreducibility, which is equivalent to requiring that the graph is strongly connected (again [Brandes and Erlebach, 2005]), or in our undirected setting, that  $G_{\mathbb{F}}$  is connected.

Jäschke et al. [2008] presented a version of *FolkRank* that circumvents the convergence issue by choosing the same  $d < 1$  for both iterations (*adapted PageRank* and baseline) and setting  $\vec{p}$  in the baseline to a vector  $\vec{p}^c$  with the same entry for any entity of the folksonomy. The same was proposed by Kim and El Saddik [2011], who furthermore observed that in this version, *FolkRank* can be computed in the single iteration

$$\vec{w} \leftarrow dA^T\vec{w} + (1-d)(\vec{p} - \vec{p}^c) \quad (8.4)$$

and is, thus, a personalized *PageRank* on the *FolkRank* graph.

Regarding the choice of the preference, Jäschke et al. [2008] described for the tag recommender use case the setting mentioned above:  $\vec{p}_u = 1 + |U|$ ,  $\vec{p}_r = 1 + |R|$ , and  $\vec{p}_i = 1$  for any other entity  $i \neq u, r$ ; while in the baseline  $\vec{p}^c$  is a vector with each entry equal to one. Both vectors are then normalized such that their entries sum up to one. The latter might have been missed by Kim and El Saddik [2011], who wrongly concluded that  $(\vec{p} - \vec{p}^c)$  in Equation 8.4 would simplify to a vector of zeroes except for the entries of the user  $u$  and the resource  $r$  which are  $|U|$  and  $|R|$ , respectively. In fact,  $(\vec{p} - \vec{p}^c)$ , taking into account the normalization, contains mostly negative values. Particularly, we have

$$(\vec{p} - \vec{p}^c)_i = \frac{1}{|V| \cdot (|U| + |R| + |V|)} \begin{cases} |V| \cdot (|U| - 1) + |T| & \text{if } i = u \\ |V| \cdot (|R| - 1) + |T| & \text{if } i = r \\ -(|U| + |R|) & \text{otherwise.} \end{cases}$$

### Item Recommendation

In the use case of item recommendation, the goal is to suggest items, given a user  $u$ . Thus, the only active entity to be used in the preference vector  $\vec{p}$  is  $u$ . Following the suggestion of Jäschke et al. [2008], mentioned above, we set  $\vec{p}_u$  to  $|U| + 1$ , all other entries of  $\vec{p}$  to one, and then normalize  $\vec{p}$  to sum up to one in total.

#### 8.2.4 Extending FolkRank

In this chapter, we explore different ways to augment *FolkRank* with further data. The *FolkRank* iteration (Equation 8.3 or Equation 8.4) has two components: the adjacency

<sup>4</sup>In that case for each entity, there are paths of length two and three beginning and ending at that entity; for example for a user  $u$  in a tag assignment  $(u, t, r)$ , the paths  $u \rightarrow t \rightarrow u$  and  $u \rightarrow t \rightarrow r \rightarrow u$ . Thus the greatest common divisor of the length of all periods of the Markov chain is one.

<sup>5</sup>Since resources and tags are contributed by users through posting, isolated nodes can only be inactive users who never created a single post.

matrix  $A$  (representing the folksonomy graph) and the preference vector  $\vec{p}$ . In the following, attending to Research Question RQ1, we present modifications to both components, that is, two options to include further data into *FolkRank*.

### FolkRank on an Extended Folksonomy

We manipulate of the underlying *FolkRank* graph  $G_{\mathbb{F}}$  (and thus  $A$  in Equation 8.3) by including another dimension  $M$ . The new structure, denoted  $\mathbb{F} + M := (U, T, R, M, Y')$ , extends the folksonomy  $\mathbb{F}$  such that  $Y'$  is a relation  $Y' \subseteq U \times T \times R \times M$ . Each triple of  $Y$  is extended with those elements of  $M$  that one of the elements of the triple is associated with. For instance, if  $M$  is a set of user groups and a user  $u$  is member of two groups  $g$  and  $h$ , then each triple  $(u, t, r) \in Y$  is extended into two quadruples  $(u, t, r, g)$  and  $(u, t, r, h)$ . If the new dimension  $M$  is the set of publication venues, then each triple  $(u, t, r) \in Y$  yields a quadruple  $(u, t, r, v(r))$ , with  $v(r)$  being the venue of the publication  $r$ . Every time a triple has no corresponding element in  $M$  (e.g., missing metadata fields), we insert a new artificial element into that triple and thus into  $M$ . The new element will be almost isolated in the graph of  $\mathbb{F} + M$ , and, thus, be of little influence. The adaptation of the *FolkRank* graph to the extended folksonomy is straightforward: Each quadruple  $(u, t, r, m)$  gives rise to six edges:  $\{u, t\}$ ,  $\{u, r\}$ ,  $\{r, t\}$ , as before, plus  $\{u, m\}$ ,  $\{t, m\}$  and  $\{r, m\}$ .

### FolkRank with Extended Personalization

The second way of including further information is the manipulation of the preference vector  $\vec{p}$ . We simply select users, tags, or resources that, next to the active user  $u$ , should receive higher preference weights and assign appropriate values to their entries in  $\vec{p}$ .

## 8.3 Related Work

In this chapter, we approach the task of improving scholarly publication recommendations in tagging systems by including further data into *FolkRank*. In the following, we review previous work grouped by the four aspects of this task: recommending scholarly publications (Section 8.3.1), resource recommendations in tagging systems (Section 8.3.2), exploiting metadata for recommendations (Section 8.3.3), and improvements of *FolkRank* through additional data (Section 8.3.4).

### 8.3.1 Recommending Scholarly Publications

The recommendation of scholarly research articles in general has been approached in a variety of studies. The proposed solutions vary depending on the available information and on the system in which the recommendations are provided. Since in this work, we focus on recommendations in social bookmarking systems, we only list a few studies of



publication recommendations in other types of systems. We refer to Beel et al. [2015] for further examples (see below).

*TechLens*<sup>+</sup> [Torres et al., 2004] is a recommendation system for digital libraries. It employs recommendation strategies based on similar references in publications and based on similarity measures computed on full texts of articles and their referenced publications. Pohl et al. [2007] used co-downloads – two documents are related through a co-download when they have both been downloaded by the same user – to recommend related scholarly articles on the preprint server arXiv. They found this usage-based information to yield more accurate predictions than co-citation and also to be more often available (many publications had few or zero co-citations). More recently, Kern et al. [2014] investigated different forms of relatedness between publications – publications sharing a venue, an author, or publications occurring in the same user profiles or user groups – by applying recommendation algorithms to suitable datasets from Mendeley. Here, especially for producing items from the same user libraries, content based recommendation using as much metadata as possible (title, abstract, authors, tags) was most successful.

A survey of recommendation systems for research papers was only recently published by Beel et al. [2015]. Discussed are more than 200 papers with respect to their approaches and their evaluation protocols, as well as possible explanations for shortcomings of several approaches.

### 8.3.2 Folksonomic Resource Recommendation

The scenario of this chapter’s investigation is the recommendation of scholarly publications as resources in a social bookmarking system. Therefore, we review several examples from literature, where publications (and sometimes other kinds of resources as well) have been recommended using data from social bookmarking systems.

Parra and Brusilovsky [2009], for instance, conducted a user study, comprising seven users and data from CiteULike, and applied 3-point scales for relevance and novelty (e.g., *relevant*, *somewhat relevant*, and *not relevant*) and hence use normalized discounted cumulative gain (nDCG), which is particularly designed for this kind of scale, as evaluation measure. Furthermore, the precision in the top  $k$  recommended items was measured for fixed numbers  $k$ . In the proposed extension of *collaborative filtering* [Sarwar et al., 2001], the results of the algorithm are additionally weighted according to the number of raters of an item. Pearson’s correlation coefficient was used as similarity measure and the results were compared to regular *collaborative filtering*, as well as to *BM25* [Manning et al., 2008].

A folksonomic resource recommendation using probabilistic latent semantic analysis has been proposed by Wetzker et al. [2009]. There, the relationships between users and resources and between resources and tags are modeled using probability distributions over latent topics and then combined in a linear combination with a weighting parameter  $\alpha$  to control the influence of both distributions. Using a Delicious dataset it is shown that using the weighted combination yields stronger precision in the

resulting recommendations than using only either the user-resource relationships or the tag-resource relationships.

Cantador et al. [2011] applied several tag similarity measures to build tag context vectors for users and items which they in turn used for item recommendation on Delicious. As evaluation measures, they used precision/recall at  $k$ , MAP, and nDCG. The best results on a Delicious dataset were achieved using *BM25*. Similarly to their approach, in this chapter, we employ tag similarity measures to boost *FolkRank* and to create baselines with *collaborative filtering*.

An approach making use of tag clusters to personalize recommendations was presented by Shepitsen et al. [2008]: A user is not only represented as a tag vector, but as a vector of a (personalized) set of tag clusters. The authors provided evidence that a user-specific choice of the set of clusters (compared to only one global clustering) yields better results on sparse data. Their approach was evaluated in a scenario where users request recommendations for a particular tag, using datasets from Delicious and last.fm. A similar approach was presented by Wartena and Wibbels [2011] with the goal of producing more diverse, topic-based recommendations. Three algorithms (among them *collaborative filtering*) were employed using tags directly and using topic clusters. Using a dataset from LibraryThing, experiments revealed that the clustering step indeed improves the recommendation performance of each of the three methods and additionally enables more diverse recommendations. Inspired by their approach, we will use tag clusters to extend the *FolkRank* graph.

Gemmel et al. [2012] built a weighted linear hybrid recommender that incorporates four *collaborative filtering* variants, a recommender suggesting the most popular resources, and an approach that directly recommends resources that are similar to the user in the tag vector space. They compared the hybrid's performance to the pair-wise interaction tensor factorization approach of Rendle and Schmidt-Thieme [2010], which had previously been used for tag recommendation. The *collaborative filtering* variants were user-based – with similarities between users being computed in the resource and in the tag vector space – and item-based – with similarities between resources being computed in the user and in the tag vector space. In contrast to plain *collaborative filtering*, this kind of hybridization enables the inclusion of all three dimensions. On all six used datasets, the hybrid outperforms each of the six contributing recommenders. The user-based *collaborative filtering* approach using the resource vector space contributes considerably to the hybrid and performs better than or comparable to the other contributing recommenders on their own. In contrast to our approach no additional metadata is included. We can repeat the observation that for user-based *collaborative filtering* the user similarities in the resource vector space work better than those in the tag vector space.

Similarly to our inclusion of group information into FolkRank, Lee and Brusilovsky [2010] incorporated information about the user's groups into *collaborative filtering* using mixed hybridization. They combine user-based *collaborative filtering* with (Jaccard) similarity, measured in the resource space, with recommendations from the group information, which in turn are a fusion of recommendations based on the group's

documents and on the group members' documents. Like in [Gemmell et al., 2012], the hybrid outperforms all the baseline approaches, which is demonstrated using data from CiteULike.

More recently, Lacic et al. [2014] proposed an item recommender based on findings from human memory theory: Candidate items are generated like in user-based *collaborative filtering* and then ranked by a combination of the similarity between the active user's previously posted items and a factor that models tag frequency and recency of these items' tags. In experiments on data from BibSonomy, CiteULike, and Movielens, they find their method to be superior to regular *collaborative filtering* and comparable or better than two other algorithms that also model the process of forgetting and remembering tags.

### 8.3.3 Exploiting Metadata for Recommendations

*FolkRank* is a graph-based method that exploits the folksonomy graph and thus the relations between users, tags, and resources, rather than information about the resources themselves. In this chapter, we experiment with content features (the publications' metadata) to enrich *FolkRank*. Therefore, here, we review literature where similar combinations of collaborative and content-based recommendations have been proposed.

An example for the benefit of metadata of web pages in tag recommendations is given by Musto et al. [2010]: The proposed algorithm generated three sets of candidates – content-based, personal, and social tags. Content-based tags are extracted from the URL, the title and the meta tags of the web page that is to be bookmarked. To choose the actually recommended tags, the three sets are ordered and tags of one set are recommended if the previous tag sets in that order are empty. Using data from BibSonomy and precision, recall, as well as the  $F1$ -measure, it is found that the combination in the above order is most successful.

Bogers [2009] presented a comprehensive evaluation of a variety of recommendation algorithms on four different datasets (from the bookmarking systems BibSonomy, Delicious, and CiteULike), and investigated the inclusion of metadata to “aid the recommendation process” as well as different hybridizations. Among the chosen algorithms, *collaborative filtering* occurred in several variations. It was found that on different datasets, different ways of including metadata produced the best recommendations but overall, the combined inclusion of different kinds of metadata that are directly associated to the publications (authors, title, journal, etc.) usually produced good results.

We complement this analysis of Bogers [2009] by evaluating *FolkRank* on similar datasets and pointing out ways to aid also this algorithm with metadata as well as social data (user groups) or usage data (recent posts of the active user). Bogers compares algorithms mainly using MAP and we follow this example.

Guan et al. [2010] approached the task of resource recommendation by mapping the folksonomic entities into a  $k$ -dimensional semantic space in which similar entities

are close together. The space is build using both the folksonomy structure as well as pairwise similarity between resources. The latter are found in the full texts of web pages on Delicious or in the abstracts of scholarly publications on CiteULike. The approach successfully outperforms several baselines, among them *user-based collaborative filtering*.

While the above mentioned work, as well as the experiments in this chapter, use metadata of resources in a social bookmarking system to improve the quality of folksonomic resource recommendations, it is also possible to go the other way around and to use the folksonomy structure to recommend entities found in the metadata. An example for the latter is provided by Heck et al. [2011], who used CiteULike data to recommend authors for scholarly cooperation. These authors are found in the metadata of the publications and are described through the sets of users who bookmarked their publications or through the sets of tags that have been assigned to their work. In a small user study comprising six physicists, it turned out that recommendations using social bookmarking data “may complement co-citation and bibliographic coupling” [Heck et al., 2011] for the purpose of recommending relevant collaborators.

### 8.3.4 Improving FolkRank by Including Additional Data

We have already presented some literature on the *FolkRank* algorithm in the previous section. Since the main contribution of this chapter is an investigation of different ways to include further data into *FolkRank*, we review other work that has approached this task.

In [Landia et al., 2012], we used the scenario of tag recommendation and BibSonomy datasets to test different modifications of *FolkRank*. Here, resources (web pages or publications) where replaced by the words they contain, specifically with words from the URL, title, and description of a bookmark and words from a publication’s metadata. Thus one resource node in the *FolkRank* graph is replaced by several word nodes. The weights of edges between words and other entities are computed using tf-idf scores. It was found that using only the words of the title to represent each resource yielded stronger results than using other metadata as well and also outperformed plain *FolkRank*. However, experiments also revealed that a simpler algorithm based on popularity of tags per word and tags per user provided even stronger results than the *FolkRank* with title words, while being computationally much less expensive.

Ramezani et al. [2010] also experimented with the weights of the edges in the *FolkRank* graph. They argued that the weight of an edge from an entity  $a$  to an entity  $b$  should not only reflect the number of times  $a$  and  $b$  occur together in a tag assignment, but also depend on the popularity of the edge’s target  $b$ . In their experiments, including popularity information into the edge weights had a strictly positive effect on the performance, measured in terms of recall and precision.

Gemmell et al. [2009] did not directly modify *FolkRank* but constructed linear hybrids, combining *FolkRank* with other recommendation strategies for tag recommendation.

On datasets from BibSonomy, CiteULike and Delicious, a hybrid with *item-based collaborative filtering* provided the best results, while combinations with other strategies like *user-based collaborative filtering*, *most popular tags by resource*, or *most popular tags by user* did not significantly improve the results or even had a negative impact.

Abel et al. [2009] presented *GFolkRank* and *CFolkRank* for the system GroupMe! In this tagging system, users can add their resources to groups, which can themselves be tagged and also be added to other groups. *GFolkRank* treats these groups as additional tags, while *CFolkRank* replaces the tag dimension by a set of tag-group pairs, where each pair with a tag  $t$  and a group  $g$  indicates that  $t$  was used in  $g$ . For both algorithms (and plain *FolkRank* as well), further modifications are applied, that (i) propagate the tags of groups to resources within the group and even tags of resources in a group to all resources in the same group, and (ii) that use the groups to set values in the preference vector  $\vec{p}$  for tags of that group. It turns out that the modified versions of *FolkRank* perform better than plain *FolkRank* when used for ranking search results of queries by tag. In a tag recommendation scenario, *FolkRank* and several modifications perform comparably well, switching positions in the benchmarking depending on the applied evaluation metric.

Similarly, Abel et al. [2010] created three versions of *FolkRank* to rank search results on TagMe!, a tagging front-end for the picture bookmarking system Flickr. All three versions made use of information rather specific to TagMe!: (i) User-added categories were added as another dimension to the *FolkRank* graph, however, categories were only connected to resources and tags but not to users. (ii) In TagMe!, tags are associated to a particular area of an image and the size of that area as well as its distance to the center of the picture were used to re-weight the edges of the *FolkRank* graph. (iii) The tags in TagMe! are automatically assigned to a fitting DBpedia URI. In the third version of *FolkRank*, the tagging dimension is replaced by the set of URIs. In a user study, the enhanced versions of *FolkRank* slightly outperformed its plain version, while a hybrid that computes the average scores of the four *FolkRank* had the highest precision results.

These previous findings suggest that modifying *FolkRank* to include further data can increase the algorithm's performance. In contrast to the experiments above, instead of modifying edge weights or using *FolkRank* in hybrids with other algorithms, our variations of *FolkRank* extend the *FolkRank* graph directly with a new component or modify the preference vector. These variations can similarly be adapted to other folksonomies where some metadata exists for the resources. With the exception of considering user groups in *FolkRank*, all modifications are independent of any further structural information like in the case of TagMe! or GroupMe! (see above).

Furthermore, all mentioned experiments used *FolkRank* for tag recommendation or for (unpersonalized) ranking of search results. As a crucial next step, we therefore evaluate several options for the incorporation of metadata into *FolkRank* to boost its performance for personalized resource recommendation.

## 8.4 Experimental Evaluation

In this section, we discuss the setup of the experiments: the applied algorithms and their parametrization, the datasets and their set-cores, and the evaluation method.

### 8.4.1 Algorithms

Here, we introduce the different versions of *FolkRank*, using the extensions described in Section 8.2.4, as well as the baselines to which we will compare *FolkRank*.

#### Extended FolkRank

We use *FolkRank* in the version that was presented by Jäschke et al. [2008], which is the one circumventing the convergence issues we discussed in Section 8.2.3. We also adapt the parameter setting  $d = 0.7$  from Jäschke et al. [2008], the same that we used in Chapter 7.

In several experiments, we add further data as new dimension  $M$  to the plain folksonomy  $\mathbb{F}$ , like described in Section 8.2.4. We denote this new structure by  $\mathbb{F}+M$ . As in BibSonomy, users are required to specify for each publication (besides the title) its authors and its year of publication, these were considered as additional dimension: The extended folksonomy with publication years is denoted by  $(\mathbb{F} + \text{publication year})$ . In the author dimension, we use either the first authors, the last authors, or all authors (and editors, if no authors are given). Author names are either normalized to their first name's initial plus lastname or to only their lastname. The according data structures are  $(\mathbb{F} + \text{authors})$ ,  $(\mathbb{F} + \text{authors lastname})$ ,  $(\mathbb{F} + \text{first authors})$ , and so on. One of the most often filled fields of publication posts are the booktitles of proceedings and the journal for articles, and we use them combined as the “venue” of a publication  $(\mathbb{F} + \text{venue})$ . Available for all posts is also the year a resource was posted, resulting in  $(\mathbb{F} + \text{posting year})$ . Choosing the venue and author dimensions is based on the rationale that usually a journal/conference or an author is focussed on a specific subdiscipline of a larger field of science and a researcher who is interested in one article of that area might be interested in the other ones from the same area, too. Selecting the years reflects the idea that often a certain topic is investigated heavily by several researchers during a (short) period of time, and thus contemporary articles might be related.

We exploit social ties among users by including the groups that some are members of  $(\mathbb{F} + \text{group})$ , usually combining users with similar interests (e.g., from the same institute).<sup>6</sup>

Finally, we make use of the semantic structure among the tags to create sets of similar tags. For that purpose, we calculate co-occurrence-based similarities between tags following the procedure described by Markines et al. [2009], and we create a graph where each tag is connected to its most similar tag. We then assign to each tag its weakly connected component in that graph as additional metadata  $(\mathbb{F} + \text{similar}$

---

<sup>6</sup>For both datasets we use the group memberships of 2012 as the older ones are not available.

tags). In a variation of this scenario ( $\mathbb{F}^*$  + similar tags), we completely omit the tag dimension from the folksonomy and replace it by the dimension of the tag-graph’s components.

### Baselines

A very simple baseline is *most popular resources*, which is an unpersonalized algorithm that simply suggests the most popular resources to any user. As our second baseline, we select user-based *collaborative filtering* (cf. Section 2.4.3), which recommends new resources to an active user based on the preference of like-minded users. We represent users both in the resource and in the tag vector space and refer to the former as  $CF_R$  and to the latter as  $CF_T$ . Furthermore, we experiment with Euclidean, Manhattan, and Cosine similarity, and we compute similarities both in the Boolean and in the non-Boolean versions of the vectorspaces. For the computation, we use the implementation in Apache Mahout.<sup>7</sup>

### 8.4.2 Datasets

The datasets that we use for our evaluation are based on the regular dumps of the publicly available data of BibSonomy. We use two datasets: With  $D_{08}$ , we denote the one from the ECML PKDD Discovery Challenge 2008 (called “rsdc08train” on the data dump web page<sup>8</sup>), which was also used by Bogers [2009]. Further, we use a larger, more recent dataset from January 1, 2012, which we refer to by  $D_{12}$ . The generation of the dataset dumps is described in [Jäschke et al., 2012], including a more in-depth description of the data from 2008.

For our analysis, we only use the publication references and ignore the bookmarks as we are especially interested in recommending scientific articles. Following the results from Chapter 7, we restrict each of the two datasets to two subsets using post-set-cores (see Section 7.2.3): For each dataset, we construct its post-set-cores at levels  $l = (0, 0, 2)$  and  $l = (20, 0, 2)$ . The resulting cores at level  $l = (0, 0, 2)$  are called  $D_{12}^R$  and  $D_{08}^R$ , and they have the property that each resource occurs in at least two posts. Thus, when we apply *LeaveXPostsOut* (see Section 2.4.2) during the evaluation, the resource of a left-out post still occurs at least once in the dataset and thus can still be selected for recommendation by any of the algorithms. Using the level  $l = (20, 0, 2)$ , we create even smaller datasets  $D_{12}^{UR}$ , and  $D_{08}^{UR}$ , in which, additionally, each user has at least 20 resources in their collection. Thus, we exclude users with only a short usage history. The sizes of the datasets and their post-set-cores can be found in Table 8.1.

### 8.4.3 Evaluation Methodology

Since it is difficult to get information on the relevance of recommendations from the users themselves, we treat their history of posted publications as gold-standard in an

<sup>7</sup><https://mahout.apache.org/>

<sup>8</sup><http://www.kde.cs.uni-kassel.de/bibsonomy/dumps/>

Table 8.1: The BibSonomy datasets and their set-cores, their sizes, and the number of test users.

dataset	users	resources	posts	tags	test
$D_{12}$	5,132	483,945	543,890	149,034	–
$D_{12}^R$	2,886	29,921	84,176	28,011	590
$D_{12}^{UR}$	541	25,072	70,382	19,998	541
$D_{08}$	1,211	71,705	92,545	28,023	–
$D_{08}^R$	729	13,001	32,962	7,084	165
$D_{08}^{UR}$	150	11,689	29,057	4,652	150

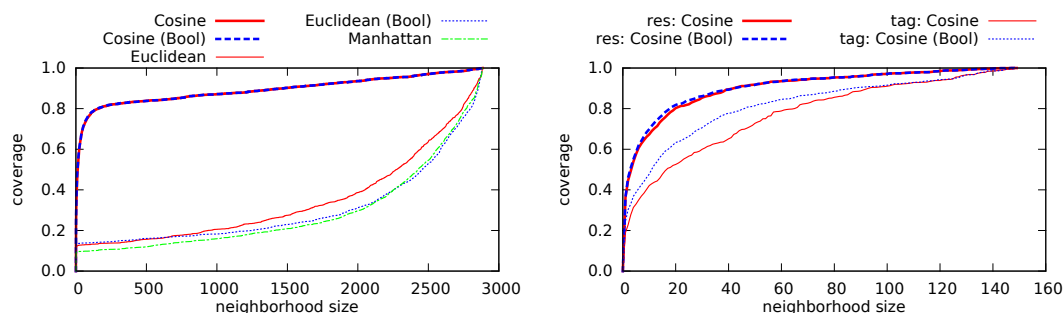
offline evaluation. Our setup leans on that of Bogers [2009], who conducted a large study with various item recommender algorithms for folksonomies, but deviates in a few subtleties.

We evaluate the recommender algorithms in the *LeaveXPostsOut* setup (cf. Section 2.4.2), using mean average precision (MAP, cf. Section 2.4.2) as quality score function. In each step of *LeaveXPostsOut*, we withhold ten posts of the active user. To avoid the *cold start problem*, we select a subset of users with large enough collections for *LeaveXPostsOut*; particularly, all users with at least 20 posts. For others, only very little is known about their interests (especially after removing ten posts). The resulting numbers of test users are shown in the last column of Table 8.1. Note that in the datasets  $D_{12}^{UR}$  and  $D_{08}^{UR}$ , every user is considered in the evaluation since they are constructed as post-set-cores with the property that every user has at least 20 posts.

Furthermore, we always leave out the ten most recent posts of the active user, rather than withholding randomly selected posts, like Bogers [2009]. This deviation has two advantages: (i) The setup is closer to the real application since the order, in which users have added their posts, is respected. (ii) Only this setup allows us to investigate the influence of recent interests, which we do in Section 8.5.4.

The application of *LeaveXPostsOut* is itself another deviation from the setup of Bogers [2009]. There, the dataset is split into a test and a training set by arbitrarily selecting 10% of the users (i.e., 15 users) as test users. Then for each such test user, Bogers selects ten arbitrarily chosen resources for testing. The remaining users and the remaining posts of the chosen users form the training set. While parameters of the evaluated algorithms are optimized using a ten-fold cross validation on the training set, the final evaluation of an algorithm's performance is conducted only on the one test set. In our experiments, we found that different selections of 10% of the set of users  $U$  (as test users) yield strong fluctuations in the resulting MAP scores due to the rather small size of the test dataset. Thus, the final result is highly dependent on the choice of the test set.





(a) Cosine similarity and similarities based on Manhattan and Euclidean distance, in Boolean and non-Boolean user profiles in the resource vector space of  $D_{12}^R$ . (The lines of Cosine and Cosine (Bool) are almost identical.) (b) All four variants of Cosine similarity (Boolean versus non-Boolean and resource vector space (res) versus tag vector space (tag)) on  $D_{08}^{UR}$ .

Figure 8.1: For various similarity functions, shown is the coverage of the ten withheld resources in the collections of the most similar users, averaged over the set of all test users  $u$ , depending on the chosen neighborhood size.

## 8.5 Results

In this section, we present our results regarding the performance of the different versions of *FolkRank* in comparison to the baseline algorithms. Before we begin with the evaluation of the algorithms themselves, we first discuss different user similarities.

### 8.5.1 User Similarities

The *collaborative filtering* algorithm, as well as the extensions of *FolkRank* we evaluate in 8.5.3, are based on the rationale that users that are (somehow) similar to the active user, are valuable sources to find resources for recommendation. Therefore, we investigate for different well-known similarity functions, how many of the most similar users it takes on average, to find many of the ten most recent items of a user  $u$  – exactly those items which we will try to recommend in the following experiments – within their collections (i.e., to yield a high coverage of these items in  $u$ 's neighborhood).

We test the Cosine similarity, as well as similarities based on Manhattan and on Euclidean distance. All of them are applied to representations of the user-profiles as resource vectors and as tag vectors. We also distinguish between Boolean representations (a user has a resource/tag at least once or not at all) and non-Boolean vectors. Note that in the Boolean case, the ordered lists of similar users according to Euclidean and Manhattan distance are identical. Also, since we withhold ten items for each considered user, a hypothetical “perfect” similarity measure would only require neighborhoods of at most ten similar users to cover all ten withheld items – which is an (extreme) upper bound for achieving coverage with as few similar users as possible.

Table 8.2: The smallest neighborhood sizes to yield a given minimum level of average coverage of the left-out resources. Displayed are the results for Cosine similarity on all datasets based on Boolean resource (res) or tag (tag) vectors.

coverage of $X_u$ in %	$D_{12}^R$ (2,886 users)		$D_{08}^R$ (1,211 users)		$D_{12}^{UR}$ (541 users)		$D_{08}^{UR}$ (150 users)	
	res	tag	res	tag	res	tag	res	tag
30	3	11	2	6	1	4	1	2
50	12	114	7	38	5	30	3	11
60	24	230	14	69	10	54	6	18
80	154	631	47	165	54	174	18	46
90	1,473	1,104	222	280	173	300	43	87

Figure 8.1(a) exemplarily shows the resulting coverage curves with different similarity functions for the largest considered dataset,  $D_{12}^R$ , using resource vectors to represent users. We can see that using Cosine similarity generates higher coverage in smaller neighborhoods than the other similarities. The results on the other three cores and those using the tag vector space are qualitatively similar.

Figure 8.1(b) shows the best performing similarities (the four variants with Cosine similarity) for the smallest dataset  $D_{08}^{UR}$ . The fraction of covered resources rises quickly to approximately 80 % (60 %) for the resource (tag) vector space. Adding further users then yields smaller gains in coverage until finally the neighborhoods containing all other users have complete coverage – as a consequence of the dataset construction each resource occurs in at least two user profiles. Using the resource vector representations of users, the coverage rises faster than when the Cosine similarity is computed on the tag vector space. Further, there is almost no difference between Boolean and non-Boolean representation, but in all cases the Boolean versions of the Cosine similarity yield comparable or slightly higher coverage especially for the smaller neighborhoods. Again, these results are similar on the other datasets.

A comparison between the four datasets is shown in Table 8.2, where neighborhood sizes for five coverage levels for the Cosine similarity are shown for both Boolean vector spaces. The numbers confirm for all four datasets that the coverage of left-out resources is higher when users are represented by their bookmarked publications. We also note that the number of similar users it takes to get a certain amount of average coverage is higher on the larger cores.

Following these observations, we will use the Cosine similarity in the recommendation experiments.

### 8.5.2 FolkRank on an Extended Folksonomy

In the following, we present the resulting MAP scores for our algorithms in different parametrizations. We start with an evaluation of differently parametrized versions of the  $CF_R$  and  $CF_T$  variants of *collaborative filtering* and *FolkRank* on an extended folksonomy. Thus, we attend to Research Question RQ2 with the first option of modifying *FolkRank*. For  $CF_R$  and  $CF_T$  we selected – according to the results in Section 8.5.1 – the Cosine similarity measure, and we experimented with different neighborhood sizes. *FolkRank* was evaluated in its original version (*FolkRank*  $\mathbb{F}$ ) and making use of further (social, semantic, or metadata) dimensions  $M$ , as described in Section 8.2 (*FolkRank*  $\mathbb{F} + M$ ). The results of these experiments are listed in Table 8.3.

As can be seen,  $CF_R$  and *FolkRank* yield comparable results. Both have much higher MAP scores than *adapted PageRank* in all experiments. Therefore, only the regular *adapted PageRank* is reported in the table and we omit results of *adapted PageRank* on extended versions of the folksonomy. Both algorithms also yield much better scores than  $CF_T$ , and finally all algorithms have higher MAP scores than the *most popular resources* baseline. Further, regular *FolkRank* ( $\mathbb{F}$ ) performs better than the different versions on extended folksonomies. The worst scores result from including the posting year or the publication year. Since only few posting years (BibSonomy started in 2006) can occur in the dataset and users tend to post publications that appeared recently, these dimensions consist of only few nodes. The induced connections between nodes of the other dimensions seem to be not meaningful for the recommendation scenario at hand. Including the venues works slightly better and we can speculate that this is due to the higher number of available venues compared to the number of years.

We can further observe that on each dataset the combinations with normalized author names yield better scores than the same version with only the authors’ last names. Again this might be due to nodes of the additional dimension connecting too many nodes in the regular three dimensions, since identifying authors only by their last name is a relatively coarse mapping. Combining  $\mathbb{F}$  with only the first authors is better than with the last authors and both are better than combining  $\mathbb{F}$  with all authors. Often, the first author of a publication is the one contributing most, and the last author often is a supervisor or department head of the other authors. It therefore seems intuitive that publications of the same first author are more interesting to a user than publications which only have arbitrary authors in common.

The inclusion of the social feature “user groups” yields the second best *FolkRank* results on the two older datasets,  $D_{08}^R$  and  $D_{08}^{UR}$ . To find an explanation here would require an in-depth study of the groups structure in the different datasets which is beyond the scope of this investigation. Finally, replacing the tag dimension in ( $\mathbb{F}^* + \text{similar tags}$ ) is slightly better than adding components of similar tags as a fourth dimension ( $\mathbb{F} + \text{similar tags}$ ).

In comparison to the results of Bogers [2009] – who also used the  $D_{08}$  dataset and excluded users with less than 20 posts, as well as publications posted by less than two

Table 8.3: MAP scores of the different algorithms in different variations, evaluated on the four datasets. Highlighted are the highest scores of each block in the table.

algorithm / variant	$D_{12}^R$	$D_{08}^R$	$D_{12}^{UR}$	$D_{08}^{UR}$
<i>most popular resources</i>	0.006	0.013	0.007	0.013
<i>collaborative filtering CF<sub>R</sub></i>				
$k = 4$	0.110	0.139	0.115	0.141
$k = 5$	0.110	0.138	0.115	0.140
$k = 10$	0.109	<b>0.141</b>	0.120	<b>0.152</b>
$k = 100$	0.112	0.130	0.116	0.139
$k =  U  - 1$	<b>0.114</b>	0.136	<b>0.121</b>	0.140
<i>collaborative filtering CF<sub>T</sub></i>				
$k = 4$	0.062	0.081	<b>0.060</b>	<b>0.088</b>
$k = 5$	0.062	<b>0.081</b>	0.060	0.081
$k = 10$	<b>0.063</b>	0.073	0.058	0.076
$k = 100$	0.051	0.055	0.052	0.057
$k =  U  - 1$	0.049	0.060	0.054	0.065
<i>adapted PageRank</i>	0.020	0.016	0.021	0.021
<i>FolkRank</i>				
$\mathbb{F}$	<b>0.111</b>	<b>0.148</b>	<b>0.123</b>	<b>0.164</b>
$\mathbb{F} + \text{authors}$	0.106	0.127	0.121	0.152
$\mathbb{F} + \text{authors lastname}$	0.097	0.118	0.109	0.141
$\mathbb{F} + \text{first authors}$	0.107	0.138	0.121	0.155
$\mathbb{F} + \text{first authors lastname}$	0.091	0.119	0.104	0.137
$\mathbb{F} + \text{last authors}$	0.104	0.127	0.115	0.146
$\mathbb{F} + \text{last authors lastname}$	0.087	0.115	0.096	0.133
$\mathbb{F} + \text{posting year}$	0.085	0.108	0.088	0.122
$\mathbb{F} + \text{publication year}$	0.089	0.099	0.095	0.111
$\mathbb{F} + \text{venue}$	0.094	0.114	0.104	0.139
$\mathbb{F}^* + \text{similar tags}$	0.099	0.136	0.110	0.148
$\mathbb{F} + \text{similar tags}$	0.097	0.126	0.108	0.145
$\mathbb{F} + \text{group}$	0.099	0.141	0.111	0.158

Table 8.4: MAP scores for regular *FolkRank* ( $\vec{p}$ ) and for two versions of *FolkRank* with preference manipulation. For each variant, the optimal neighborhood size  $k$  is shown next to the MAP score.

<i>FolkRank</i> variant	$D_{12}^R$		$D_{08}^R$		$D_{12}^{UR}$		$D_{08}^{UR}$	
	k	MAP	k	MAP	k	MAP	k	MAP
$\vec{p}$	–	0.111	–	0.148	–	0.123	–	0.164
$\vec{p}$ + similar users	1	0.121	1	0.152	1	0.132	1	0.168
$\vec{p}$ + recency	9	<b>0.129</b>	74	<b>0.159</b>	13	<b>0.141</b>	57	<b>0.179</b>

users –, we yield higher scores for the same algorithms (e.g., 0.141 instead of 0.087 for  $CF_R$  with neighborhoods of size 4). We conjecture that this is due to differences in the setups: our scores are based on the whole set of users instead of a random sample of only 15 users.

### 8.5.3 Exploiting Similar Users

In the previous section, we saw that regular *FolkRank* had higher scores than the versions on an extended folksonomy, thus a negative result regarding our second research question (RQ2). However, the observation that  $CF_R$  and *FolkRank* perform comparably on all datasets is a good motivation to try to incorporate similar users, which are the basis of  $CF_R$ , into *FolkRank*. We achieve this by modifying the preference vector  $\vec{p}$  of *FolkRank*, and we refer to this version by ( $\vec{p}$  + similar users) in the following. For a target user  $u$ , we select the  $k$  most similar users (according to the Cosine similarity measure<sup>9</sup> in the resource vector space) and insert their similarity to  $u$  as weight into  $\vec{p}$ . More precisely, let the set of all users in  $U \setminus \{u\}$  be ranked by their similarity to  $u$ , and let  $\text{rank}_u(v)$  be the rank of a user  $v \in U \setminus \{u\}$ . Then, we set

$$\vec{p}_i := \begin{cases} 1 + |U| & \text{if } i = u \\ 1 + |U| \cos(\vec{x}_u^R, \vec{x}_i^R) & \text{if } i \in U \text{ and } \text{rank}_u(i) \leq k \\ 1 & \text{otherwise.} \end{cases}$$

The results for different neighborhood sizes  $k$  are depicted in Figure 8.2: All scenarios (the two algorithms *FolkRank* and *adapted PageRank* on the four set-cores) profit from the inclusion of at least small neighborhoods. On each dataset, *FolkRank* achieves the best results when only the single most similar user is getting additional preference. Here, *FolkRank* without additional preference is outperformed. The exact scores are

<sup>9</sup>As expected, considering the findings in Section 8.5.1, using the Euclidean distance to construct the neighborhoods did only decrease the recommendation quality.

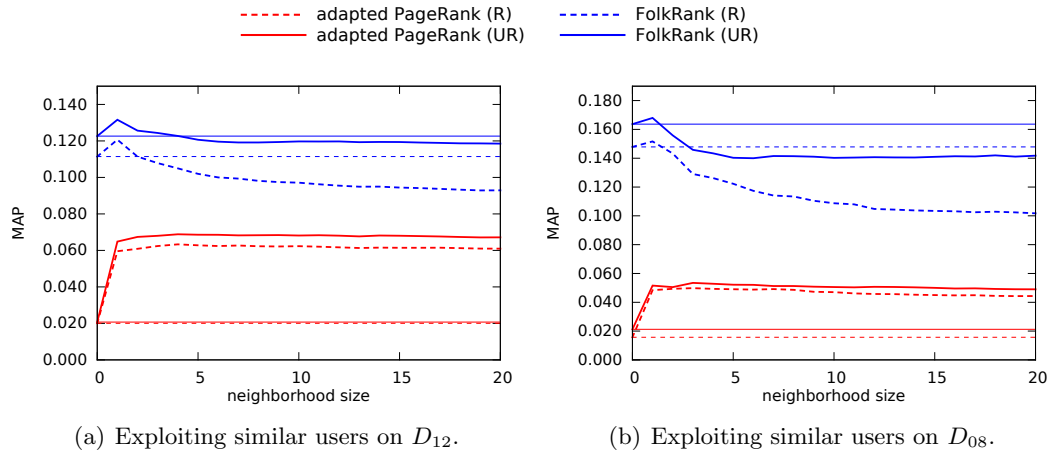


Figure 8.2: MAP scores for *FolkRank* and *adapted PageRank* with preference modification ( $\vec{p} +$  similar users). The straight lines show the according MAP score without additional preference.

reported in Table 8.4. Increasing the neighborhood size decreases the average MAP scores of *FolkRank* even below the score of the plain *FolkRank* quickly. Although *adapted PageRank* cannot compete with *FolkRank*, it is worth noting that it profits even more from the inclusion of similar users, more than tripling the average score on  $D_{12}$ . In contrast to *FolkRank*, this effect is strong also for larger neighborhoods.

To test the significance of the improvement of the modified *FolkRank* over the plain one, we employ a sign test (cf. Salzberg [1997], Demšar [2006]): For each test user, the MAP scores of both *FolkRank* versions are compared, and we count the wins and losses of the version with additional preferences ( $\vec{p} +$  similar users). With the sign test, we infer whether the number of users for which ( $\vec{p} +$  similar users) yields better results, is significantly higher than the number of users for which plain *FolkRank* wins, given a significance level  $p$ . In Table 8.5, we report the results for the four set-cores for  $p = 0.05$ . The row “threshold” marks the number of wins that ( $\vec{p} +$  similar users) would have to exceed to be considered significantly better than plain *FolkRank*.

We can observe that not only there is no significance, but actually, plain *FolkRank* wins more often than ( $\vec{p} +$  similar users). This means, that for a majority of users, the MAP scores drop, while there are few users that profit relatively well from the inclusion of similar users, causing the overall increase of the average MAP score. A consequence from this result would be, to use not only a personalizing recommender algorithm but actually to personalize the choice of the algorithm, offering recommendations from ( $\vec{p} +$  similar users) to one set of users and use regular *FolkRank* for the others. This would, however, require a method to predict for a user which algorithm will likely be more successful.

Table 8.5: Wins and losses of *FolkRank* with additional preference for similar users ( $\vec{p}$  + similar users) in comparison to regular *FolkRank*: Per user, we compare the MAP score of the manipulated *FolkRank* to the MAP score of the regular *FolkRank*, and we count the number of users for which the manipulated version produces higher (wins) or lower (losses) scores. We use the sign test to determine whether the manipulated *FolkRank* outperforms the regular one on significantly many users (significance level  $p = 0.05$ ).

$\vec{p}$ + similar users	$D_{12}^R$	$D_{08}^R$	$D_{12}^{UR}$	$D_{08}^{UR}$
wins	268	65	259	67
losses	319	99	275	80
threshold	318	95	293	87
significant	–	–	–	–

#### 8.5.4 Exploiting Recent Resources

In the next experiment, we take into account that a user’s interest may vary over time. It seems reasonable to expect that recently posted resources are an indicator for the current interests of a user. Like in the experiment with similar users in the previous section, we modify the preference vector  $\vec{p}$  of *FolkRank* by assigning the same weight to all considered recent resources: Let the set of resources  $R$  be ordered by the time at which the active user  $u$  posted them (the most recently posted resource at the first position) and let  $r_u(r)$  be the rank of resource  $r$  in that order. For resources that  $u$  has not posted, we set  $\text{rank}_u(r) = \infty$ . We set the preference vector  $\vec{p}$  to

$$\vec{p}_i := \begin{cases} 1 + |U| & \text{if } i = u \\ 1 + |R| & \text{if } i \in R \text{ and } \text{rank}_u(i) \leq k \\ 1 & \text{otherwise.} \end{cases}$$

We will refer to this version of *FolkRank* as ( $\vec{p}$  + recency). The diagrams in Figure 8.3 show the resulting MAP scores for both *FolkRank* and *adapted PageRank*, and the top values of *FolkRank* are again reported in Table 8.4. On the two more recent set-cores,  $D_{12}^R$  and  $D_{12}^{UR}$ , the scores rise immediately above the score of plain *FolkRank*, while on the datasets from 2008, they first decrease but then also exceed the baseline, using seven (five) or more recent resources on  $D_{08}^R$  ( $D_{08}^{UR}$ ). The optimal values are achieved at very different sizes. However, including larger numbers of recent resources yields quite stable results that are almost as good as the optimum. This phenomenon can in part be explained by the fact that often users do not even have that many resources to be used in  $\vec{p}$ , and therefore, the preference vector does no longer change for higher  $k$ . Again *adapted PageRank* results also improve significantly but not to the level of *FolkRank*. It seems that it is not particularly the recency but rather the inclusion of

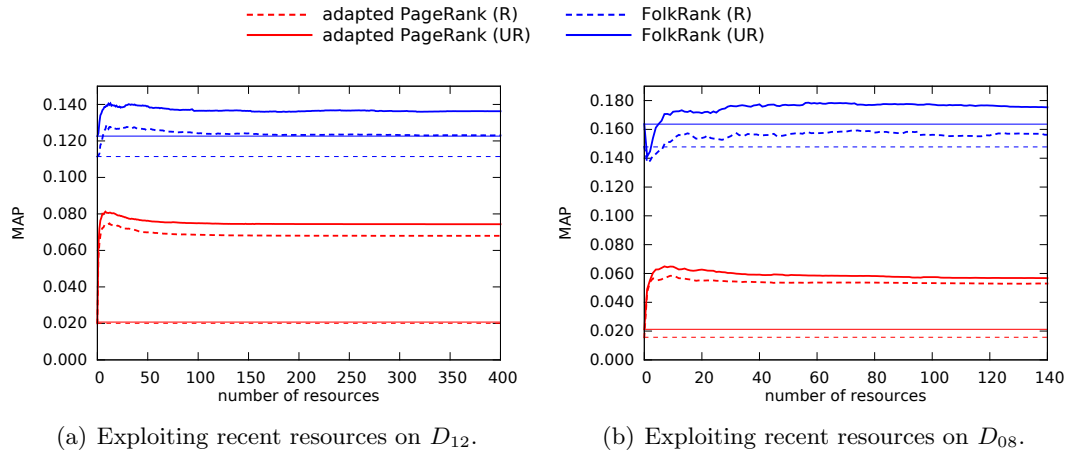


Figure 8.3: The MAP scores for *FolkRank* and *adapted PageRank* with preference modification ( $\vec{p} + \text{recency}$ ) for different numbers of included recent resources. The straight lines show the according MAP score without additional preference.

many resources a user had previously posted that has a positive effect on the *FolkRank* scores.

Like before, we employ the sign test to check whether the improvement of ( $\vec{p} + \text{recency}$ ) over plain *FolkRank* is significant. We set the number of included resources to the values in Table 8.4. The wins and losses are shown in Table 8.6. This time, on all four cores, the modification of the preference benefits many more users than it spoils. The test’s conclusion is that the improvement is significant. We can answer Research Question RQ2, by stating that the extension of *FolkRank* where additional preference is assigned to the resources a user had previously bookmarked, is the best-performing strategy in our benchmark and a significantly better approach than regular *FolkRank*. Still, we also observe that there is a large minority of users that had better scores with plain *FolkRank*, which is again an indication to use different algorithms for different users.

## 8.6 Conclusion

In this chapter, we have approached a more proactive means of supporting researchers in social bookmarking systems: recommending scholarly literature. In our experiments, we found *FolkRank* to be comparable to *collaborative filtering* when users are represented as their tagged resources. *FolkRank* outperformed *adapted PageRank*, *collaborative filtering* based on user representations in the tag vector space, as well as the *most popular resources* baseline. Our experiments yielded better results than Bogers [2009] for *collaborative filtering*, presumably due to the slightly different setup.



Table 8.6: Wins and losses of *FolkRank* with additional preference for recently posted resources ( $\vec{p}$  + recency) in comparison to regular *FolkRank*: Per user, we compare the MAP score of the manipulated *FolkRank* to the MAP score of the regular *FolkRank*, and we count the number of users for which manipulated version produces higher (wins) or lower (losses) scores. We use the sign test to determine whether the manipulated *FolkRank* outperforms the regular one on significantly many users (significance level  $p = 0.05$ ).

$\vec{p}$ + recency	$D_{12}^R$	$D_{08}^R$	$D_{12}^{UR}$	$D_{08}^{UR}$
wins	354	107	343	99
losses	233	57	191	48
threshold	318	95	293	87
significant	✓	✓	✓	✓

For the inclusion of metadata in *FolkRank* as an additional dimension, we found that it does not improve the overall recommendation performance. However, some of the additional dimensions (authors or groups) yielded comparable results. Like shown in [Bogers, 2009], different recommenders perform differently on different datasets. Hence, a reasonable next step would be to compare the more successful metadata strategies on other datasets and to investigate whether certain users can benefit more from the inclusion of certain kinds of data than others. While the idea to incorporate the “more of the same”-idea with authors or venues for scholarly publications did not pay out, it would still be worth experimenting with that same strategy on other tagging systems where the tagged resources have an even closer relationship to the added dimensions (e.g., adding bands as fourth dimension in a folksonomy where the resources are songs).

For the inclusion of similar users, we saw that small neighborhoods are suitable to improve *FolkRank*’s recommendations for a minority of users. For the selection of similar users, the Cosine similarity is the measure of choice. We also showed that the previously posted resources of a user are a valuable indicator for the current interests of a user. Including previously posted resources yielded the best results of *FolkRank* in our experiments. Although these resources are directly connected to the active user in the *FolkRank* graph, using them in the preference vector still had a significant positive impact.

By and large, regarding our first research question (RQ1), we found two ways of modifying *FolkRank*. Regarding Research Question RQ2, we saw that among all modified versions of *FolkRank* and compared with all baselines, the most successful approach to recommending scholarly publications in BibSonomy is modifying *FolkRank*’s preference vector by assigning weights to the resources a user had previously posted.

### 8.6.1 Future Research

The results of *FolkRank* with modified preference vector suggest that the success of different modifications depends on the user to whom the recommendations are provided. To tackle the task of finding the best fitting recommender algorithm per user, methods like subgroup discovery – to determine characteristic descriptions for groups of users that respond well to a particular type of recommender – or classification – to map each user to a particular recommender directly – could be applied. In Appendix E, we present a preliminary study in that direction by comparing several recommenders' performance to various aspects of user behavior in BibSonomy.

Furthermore, it would be worth investigating the performance of *FolkRank* in different parametrizations on other datasets. Particularly, parameters could be learned for the optimal inclusion of other data, like for choosing the preference weights in  $\vec{p}$  or the numbers of included similar neighbors or previously posted resources. Finally, despite the weaker performance when further dimensions are included, it might well turn out that certain combinations of the here proposed methods yield actually better results.

Truly capturing a recommender's recommendation performance requires an online evaluation with many users, since offline evaluations can only determine how well an algorithm can retrieve resources a user has already found, without the algorithm's help. Here, BibSonomy suggests itself and respective experiments are already planned, using the soon to-be released recommendation framework.<sup>10</sup>

---

<sup>10</sup><https://bitbucket.org/bibsonomy/recommender-framework> (accessed July 15, 2015)

## Chapter 9

### ◆ Opportunities and Risks of Online Literature-Reviewing Systems



Peer review has been the predominant tool for the evaluation of scholarly publications. However, it has been criticized for various reasons, among others fake reviews leading to accepts of faulty publications or invalid reviews leading to unnecessary rejects. Social peer review is a post-publication quality evaluation. During the usage phase of a publication, it is criticized online and subject to discussion between the authors and their peers. Online social peer review can have benefits for the reviewed publications, like higher visibility through the received attention. It can also have severe consequences for the authors when the criticism is negative – opposing the published results or attesting low quality or irrelevance. Therefore, in such systems, the rights of the concerned parties (referees and authors) must be treated with care. In this chapter, we discuss several requirements on the design of a social peer-reviewing system and various technical means to meet them. We first cover the case of online rating systems in general before we move to the special case of evaluating scholarly literature. We compare social peer review to other forms of publication evaluation, namely traditional (closed) and open peer review, as well as implicit evaluations through usage and citation metrics. Furthermore, we address the integration of social peer review as a secondary feature into social tagging systems.

#### 9.1 Introduction

In the creation phase of the publication life cycle, authors of a scholarly publication (or, in that phase, rather of a manuscript) select a publication venue (e.g., a conference or a journal) to which they submit it. The editors of that venue will usually employ a *peer review* process to determine whether or not to publish the manuscript. The decision is based on the opinions of selected peers who read the manuscript and provide a review of its quality and the merit of the presented results. Similarly, research proposals for project calls are evaluated by peers of the applicants to decide which proposals will receive funding.

However, the traditional form of pre-publication peer review has been questioned time and again regarding its ability to ensure quality, its fairness, and its scalability.

Spectacular cases of fraud, like the doctoral thesis of the German politician Karl-Theodor zu Guttenberg,<sup>1</sup> or retracted articles due to faked peer review in highly respectable venues [Ferguson et al., 2014] have exposed weaknesses in the peer review process. Moreover, the rising number of publications [Bornmann and Mutz, 2015] implies a rising demand for peer reviews and thus increases the burden on the potential reviewers. Finally, through pre-print servers, like the arXiv, manuscripts are available to potential readers and can be cited before they are even submitted to peer review.

The negative influences of all these issues can at least be mitigated through forms of post-publication evaluation, particularly through social peer review. This form of evaluation takes place in the usage phase of a publication and can be applied to previously unreviewed material, as well as to articles that have already passed peer review before their publication. Social peer review is a web-based form of collaborative publication evaluation that combines the function of traditional peer review with the social nature of the Web 2.0. Depending on a social peer-reviewing system's design, users can review publications in different forms (short comment or full-blown review), add numerical ratings, or discuss with authors. In that way, social peer-reviewing systems provide the tools for researchers to aid their colleagues by indicating excellent or unworthy reads. They enable a form of scholarly communication, which can enhance the visibility of publications, serve as collection of criticism (errata), and help researchers find high-quality publications independent from the reputation of their publishing venue. In contrast to pre-publication peer review, it is not bound to any time-frame. Thus, older articles can receive the same attention as younger ones, and literature that has passed peer review can be subjected to critical review time and again. A social-peer-reviewing platform can be implemented as a stand-alone tool, or it can be integrated into other publication-minded systems, such as scholarly tagging systems, like BibSonomy (see Section 2.3.2), CiteULike,<sup>2</sup> or Mendeley.<sup>3</sup>

As it is social by nature, social peer review thrives on the active participation of researchers who are willing to comment on publications or to rate them. Many social systems grant access to anyone who registers (usually requiring only an e-mail address or some other token of identification). Hence, it must be ensured that the critical comments actually constitute a peer review and not only a crowd review, meaning that the reviewers are actually qualified to assess the respective publications. Especially the rights of these publications' authors must be protected since strong and, worse, invalid criticism has the potential to severely harm a researcher's reputation and career. Moreover, personality rights must be respected, protecting both authors and reviewers, since publishing personal opinions bears the significant risk of portraying individual persons or their products (e.g., authors, or their research) in an overly positive or overly negative light. By and large, social peer-reviewing systems, like all online rating systems, must be compliant with legal requirements (see Section 2.5). It

---

<sup>1</sup>[http://de.guttenplag.wikia.com/wiki/GuttenPlag\\_Wiki](http://de.guttenplag.wikia.com/wiki/GuttenPlag_Wiki)

<sup>2</sup><http://www.citeulike.org/>

<sup>3</sup><https://www.mendeley.com/>

is the responsibility of such systems' operators to design them in a socially compatible way. Therefore in this chapter, we first discuss several aspects of arbitrary online rating systems, regarding their influence and social impact, as well as possibilities of misusing such systems to harm individuals. Afterwards, we address the particular case of rating scholarly literature in a social peer-reviewing system, which has additional risks, but which also presents an expedient addition to classic pre-publication peer review. Next to social compatibility, there are also the questions of who should run a social peer-reviewing system and where can it be integrated. We propose a system run by the research community and its integration into publication management systems.

With this chapter, we conclude the scholarly social bookmarking theme of this thesis with a discussion on opportunities and risks of social peer review. Social peer review can be implemented in scholarly social tagging systems as a feature that extends them beyond the core functionality of tagging. By commenting and rating publications, researchers can help their colleagues find the high quality publications and avoid faulty ones. Thus, implementing such a feature is a means to help researchers support other researchers.

**Research Questions.** We describe and discuss opportunities and risks that arise in online rating systems. First, we address such systems in general and then for the specific case of social peer reviewing of scholarly publications. Thus, our research questions are:

- (RQ1) Which opportunities and risks do online rating systems bear in general?
- (RQ2) Which opportunities and risks does the choice of peers in online social *peer* review bear?
- (RQ3) Which opportunities and risks does social peer review bear in comparison to other forms of evaluating the quality of publications?

**Contributions.** In this chapter, we discuss opportunities and risks of online social peer review, a means of researchers supporting each other in navigating the body of available scholarly literature.

1. We address four aspects of online rating in general, and we discuss the suitability of several features regarding their social acceptability.
2. We review risks and opportunities of various forms of publication quality evaluation, before we address them specifically for social peer review as a specialized form of online rating.
3. We propose a social peer-reviewing system, run by the research community and discuss how social peer review can become part of a scholarly tagging system.

**Limitations.** Some aspects of social peer review are to be viewed critically before the background of legal requirements. For that, we refer to our book [Doerfel et al., 2013b], where they are discussed in greater detail. Furthermore, we always refer to German law. This restriction must be made, as many countries have different laws and it is impossible to consider them all. Since all authors of the above mentioned book were German (and particularly the co-authors with judicial background were experts in German law), we naturally selected German law as the legal basis to build onto. In this chapter, we discuss many technical features, such as user authentication or verification of expertise in a web system. Obviously, considering the fast-paced developments on the web, these matters are subject to quick change. In this light, our discussion here presents a snapshot of the current state and has no claim to completeness. There may well be further ideas to realize different features or to defend social peer review against misuse, and so on. However, we contribute a view on social peer review from the perspective of social compatibility, and we list many opportunities as well as risks, which need to be considered carefully, when realizing such a system in practice.

**Structure.** To approach the topic of online social peer review, we first discuss online rating systems in general in Section 9.2 and then different forms of evaluating the quality of scholarly publications in Section 9.3. In Section 9.4, we peruse opportunities and risks of the various forms of publication quality assessment – focusing particularly on online social peer review. In Section 9.5, we address the actual implementation of a social peer-reviewing system within a scholarly social tagging system, and Section 9.6 concludes the chapter.

The arguments presented in the following have previously been published in the German-language publications [Kartal et al., 2011] and [Doerfel et al., 2013b]. This chapter is a (slightly restructured) translation where the judicial part of the two earlier publications – which had mainly been contributed by co-author Aliye Kartal-Aydemir – has been reduced to those aspects that form the basis for the more technical discussion.

## 9.2 Design Features of Online Rating Systems

In online rating systems, different kinds of resources are reviewed by the system’s users. Resources can be objects, like articles of a shop; persons, like teachers, lecturers, or doctors; or creative work, like scholarly publications or books. The resources are usually described by some metadata – a set of attributes that depends on the type of resource, often a name or title, a description, an image, or even the full resource (e.g., a publication’s full-text). Resources can be added by the system’ operators (e.g., in a web shop) or by the system’s users (researchers add publications, pupils add their teachers, etc.). The latter is, for instance, the case in a tagging system with reviewing features: Users contribute and tag resources, and thus, the corpus of available ratable entities arises from that user activity. Particularly interesting in this case is that the owners or creators of the resources (e.g., the authors whose publications are rated

or the company that produces the rated items) need not be aware of the ratings. Furthermore, allowing users to add resources opens the door for misusing the rating system by adding irrelevant or inappropriate resources (spam) or by providing false descriptions.

Often, reviewing is only a secondary feature that extends the utility which the portal provides, such as buying products (web shops) or collecting and cataloging resources (tagging systems). As Web 2.0 portals, many rating systems offer – next to the feature of reviewing – means of social interaction and networking: Users can declare friendships, follow other users, or join groups. These links between individuals can be used for visibility restrictions: Users might want to make their user name or (some details of) their review visible only to a particular group of users, but not to the whole public.

The relevant basics from the judicial point of view on online rating systems are the freedom of opinion and the freedom of information, mainly protecting the reviewers, as well as personality rights and occupational freedom of the reviewees. The right to informational self-determination protects both the system's users and the persons who (or whose products) are rated. See Section 2.5 for a brief description or [Doerfel et al., 2013b, Chapters 3 and 5] for a more extensive discussion of the judicial aspects of online ratings. A legal case that was prominently discussed in the (German) media, was that of the teacher rating portal *spickmich*.<sup>4</sup> After having received negative (anonymous) reviews, a teacher had filed suit against the portal's operators. The case was processed in several courts, before it was finally brought before the Federal Constitutional Court of Germany, where it was dismissed [Pressestelle des Bundesgerichtshofs, 2009]. In a similar case, the court dismissed the suit of a gynecologist who had demanded the deletion of data referring to him, including anonymously published ratings, in the medical doctor recommender platform *jameda*<sup>5</sup> [Pressestelle des Bundesgerichtshofs, 2014].

In the following, we discuss several design aspects of online rating systems and their opportunities and risks, thus attending to our first research question (RQ1). For operators and developers who design, maintain, and run rating systems it is imperative to be aware of the risks of particular design choices. Throughout this section, we will often assume that the rated entities (resources) are persons, like teachers, medical doctors, or scientists, who are either rated directly or indirectly through ratings of their work.

### 9.2.1 Rating in Closed User Groups

The idea of closed user groups is to restrict reading and adding ratings to those who have a justified interest in the rated resources. To be able to contribute ratings, it is usually required to register an account. Typically, users have to provide an e-mail address and – depending on the system – additional information, like the affiliation

---

<sup>4</sup><http://www.spickmich.de/> (inactive since 2014)

<sup>5</sup><http://www.jameda.de/>

with some institution (school, university, etc.). To activate the account, users are sent an e-mail with an activation link. By clicking it, users verify that they indeed have access to the e-mail address. Some systems further accept or require other authentication tokens like a mobile-phone number, which can be verified similarly. Others allow the authentication using third party services, such as a Facebook account or the OpenID protocol.<sup>6</sup> For the latter, a user registers an account with an OpenID provider and receives an authorization token (an ID). The ID can then be used to register with the rating system (or with other online services).

By requiring users to register, the circle of raters is somewhat restricted. However, usually anyone may register, provided that they volunteer their e-mail address (or the respective authentication token). Thus, it suggests itself to use an additional restriction, for instance, requiring the affiliation with a particular institution. In German law, the criterion for the adequacy of such restrictions is the consideration of a valid public interest: For instance, for rating medical doctors, like on AOK-Arztnavigator,<sup>7</sup> the circle of potential patients can hardly be narrowed down to a particular set of persons; it is rather a subject of public interest. However, school-internal ratings of teachers, like on the German portal spickmich, are suitable for an according restriction.

The technical realization of a meaningful restriction is rather difficult, especially since usually only little is known about (potential) users. A verified e-mail address is not a secure restriction since users can readily have more than one address, and they can easily create further e-mail addresses anonymously (e.g., using temporary e-mail services, like TempEMail<sup>8</sup>). Sometimes, the rating system's context suggests obvious restrictions: For example, a portal for rating university professors could allow viewing or adding ratings only to those who are enrolled at the university that the professor in question is affiliated with. An easy-to-realize technical solution would be to allow only those users who have verified an e-mail address of that university. Similar restrictions are, however, not available for other rating contexts. For example, in a scenario where pupils rate their teachers, requiring a school e-mail address would exclude all schools who do not hand out e-mail addresses to their pupils, from using the system at all.

In other contexts, already the process of defining a suitable circle of users is a problem. For example in the scenario of rating scholarly publications, a reasonable, yet vague restriction would be "scholars", or even more restrictive, "scholars with own research experience in the respective area". In such cases, a list of verifiable criteria would have to be established to select the raters. We will come back to this particular scenario in Section 9.4.1.

## 9.2.2 Mode of Rating

Different portals have offered various assessment modes, including criticism in free-text form, the selection of properties (e.g., in the teacher-rating setting of spickmich:

---

<sup>6</sup><http://openid.net/>

<sup>7</sup><http://weisse-liste.arzt.aok-arztnavi.de/>

<sup>8</sup><http://www.tempemail.net/>



“humane”, “fair”, “motivated”), or the quantified assessment of particular traits or of the resource itself, using ordinal (e.g., “does not apply”, “applies moderately”, “applies fully”) or cardinal scales (e.g., star rating).

One basic distinction between different modes of rating is whether they allow the aggregation of ratings. Aggregation is always possible, when the rating scale is cardinal (star ratings, school grades, etc.). When evaluations consist of free-text an aggregation is not directly possible. There are, however, means to mine opinions from such reviews and to create summaries for different discovered properties, like technical features of a product or character traits of a person [Hu and Liu, 2004]. A way to enforce a fix number of aspects that should be rated individually, is to provide a set of subcategories. In each such subcategory, the mode of evaluation can again be quantitative or textual. This is, for example, the case in the portal Peer Evaluation,<sup>9</sup> where scholarly publications are reviewed according to such criteria as “novelty and originality”, “methodology and results”, and “writing”.

From the perspective of legal compatibility, all above mentioned modes can bear problematic aspects, especially when the evaluated resources are persons. The judgment whether or not a particular evaluation is appropriate, must be determined in the individual case. We briefly mention a few aspects that can play a role: Through the use of explicit criteria (subcategories, features), a review can have the character of a factual claim (i.e., a claim that can be proven right or wrong). Often, however, reviews are rather the utterances of opinions or a mix of facts and opinion. These are protected under the freedom of opinion (Article 5 in German Basic Law). Furthermore, the adequacy of ratable attributes depends on the *sphere* they concern – in German law, a person’s personal traits are grouped into several spheres, which have different degrees of protection. Particularly relevant for the rating of persons (or their work) are the *social sphere* and the *private sphere* (which is stronger protected than the social sphere). General assessments, like a single star rating, bear less risks, since no particular (private) attribute of the rated person is evaluated. On the downside, such ratings offer less transparency – it is unclear what in particular is good or bad about the rated resource. To compensate, ratings can be accompanied with a free-text review (like on Amazon, BibSonomy, or CiteULike). The latter, however, bears the risk of misuse for libel, which must be avoided. One option therefor is a means to report misuse to the platform’s operator.

### 9.2.3 Aggregated Ratings

One feature that provides utility to the readers of rating systems is the aggregation of individual ratings. When particularly many raters have assessed the same resource, aggregated ratings help users grasp an overall impression. Moreover, resources can be ranked according to their aggregated ratings, in search result lists or in recommendations. However, the question arises whether the aggregations are representative

---

<sup>9</sup><http://peerevaluation.org/>

and appropriate. This aspect is problematic when through false conclusions the users' freedom of information is affected.

Following the arguments of Kamp and Peifer [2009], several aspects have to be considered regarding statistical validity: The set of received ratings (for one resource) is a sample drawn from the opinions of some greater group of users (compare Section 9.2.1). Moreover, the sample is not drawn randomly nor with respect to the coverage of demographics or psychographics. Rather, the raters volunteer to rate, following their own motivation, which distinguishes them from non-raters. According to Kamp and Peifer [2009], especially when the circle of potential raters is small, like in the scenario of spickmich, a representative sample would have to contain a large portion of the potential raters (e.g., 75 % to 80 % if the group of potential raters has 20-30 members, like school classes typically do). In most cases, it can be expected that the required minimum sample size would be so high that the probability of achieving that many ratings and the time this would take, are in no reasonable relation to the life cycle or the interestingness of the resource. Some portals require that at least some fix minimum threshold on the number of raters must be met, before an aggregated rating is displayed. However, this procedure ignores that the threshold for a representative sample does depend on the number of potential raters of the individual resource.

Another choice to be made is the aggregation function. A commonly used statistics is the arithmetic mean of the (numeric) ratings, which has the advantage that it is easily understood and verifiable by the system's users. Some portals, however, employ more complicated estimators. For instance, the movie portal IMDb<sup>10</sup> uses a Bayesian estimate, as well as an undisclosed weighting scheme for the aggregation of votes,<sup>11</sup> to compile a list of top rated movies.<sup>12</sup> Aggregation functions vary not only in the way they compute the resulting score, but also in the way individual ratings are interpreted. Different users can have a different understanding of a rating scale. Some users might prefer using only the extreme scores, others might only use the top scores on the scale, even when they dislike something (e.g., rating a bad item with "only" four out of five stars). Aggregations can therefore consider a user's previous ratings to create a new score that is comparable to the ratings of others. Further parameters to consider could be the timeliness of ratings, the experience of the raters, their acquaintance with the subject (e.g., in a scenario of rating scholarly literature their expertise in the respective field). Finally, an approach to delegate the weighting of ratings to the users, is to allow the review of reviews. In such a setting, the review of a resource can itself be reviewed, for example, on Amazon it can be marked as "helpful" or "not helpful". More helpful ratings can be included with higher weights into the overall result. While the above mentioned techniques include all ratings, another approach to handle explicitly unfair ratings, is to detect them and to ignore them in aggregated ratings. Such methods have been described, among others, by Dellarocas [2000] and Whitby et al. [2004].

---

<sup>10</sup><http://www.imdb.com/>

<sup>11</sup>[http://www.imdb.com/help/show\\_leaf?votestopfaq](http://www.imdb.com/help/show_leaf?votestopfaq)

<sup>12</sup><http://www.imdb.com/chart/top>

Submitting a particularly unfair review is only one way of manipulating an aggregate rating. Another way is to create more than one account and use them to submit many ratings. A relatively safe method to hinder registering more than one account, is requiring a unique identifier at registration. While a user can easily generate multiple e-mail addresses, it is harder to acquire other tokens, such as mobile phone numbers. An extreme example would be to require the citizen ID document (e.g., in Germany, this is possible using the postident procedure<sup>13</sup>). However, while such provisions seem appropriate for applications like online banking, they seem disproportionate for some product or service rating system. A second strategy is to increase the amount of time it takes a user to register accounts. This can be accomplished using a so-called *captcha* that has to be solved by a user upon registration. Captchas are small tasks (e.g., identifying some object in a picture) that are difficult to solve for machines, but not for humans. Captchas hinder the automated creation of accounts, however, one must consider that often resources receive only few ratings. Therefore, only few accounts must be created to drastically influence an overall rating and users might find the effort of solving a few captchas manually acceptable. Further obstacles can be implemented for hindering misuse, for instance, techniques adapted from spam protection. Often the critical factor is the timing. For example, a rating system could require a minimum membership duration before a user can submit reviews, or it might count only ratings from users who submit new ratings regularly.

### 9.2.4 Ratings in Search Engines

Rating portals can choose to make the ratings visible to the public. Moreover, these ratings can be displayed within the result lists of search engines and thus outside the context of the rating system. Special markup, like schema.org,<sup>14</sup> can be included into web pages to be interpreted by web search engines. For instance, using the schema.org entity “Rating”,<sup>15</sup> a numerical rating can be marked on a web page. Search engines can use such information to enrich the information that is displayed next to a hit – usually a small snippet of text is shown next to the link. Rich snippets can display ratings (among others) directly there. Users can perceive the rating without having to visit the respective portal.

The public visibility of ratings can be in conflict with the requirement of a closed user group (Section 9.2.1). A justified public interest must be weighted against the protection of the rated individuals. Particularly, in the case of scientists or their publications, ratings are of interest to the research community, yet negative ratings can have a severe impact on an author’s career.

---

<sup>13</sup><http://www.deutschepost.de/en/p/postident.html>

<sup>14</sup><http://schema.org/>

<sup>15</sup><http://schema.org/Rating>

### 9.2.5 Summary

The design of online rating platforms has several aspects that bear the risk of violating personality rights (and potentially others as well), and they can contain features that make (intentional) misuse for the purpose of libel possible. In the case of the German teacher-rating portal *spickmich*, the Federal Court of Justice has discussed several properties in their verdict [Pressestelle des Bundesgerichtshofs, 2009], concluding that ratings in this particular portal are acceptable, dismissing a case made by a teacher against *spickmich*. However, in the verdict it is explicitly pointed out, that the admissibility of online reviews must be scrutinized in each individual case. Operators of such rating platforms must be aware of the risks. Attending to Research Question RQ1, above, we have listed several options for designing rating platforms that can help reduce the chances of harming the rated individuals. However, fast technological progress will always be accompanied by both new opportunities and new risks that legislation will have to react to.

## 9.3 Four Models of Publication Quality Evaluation

Before we discuss opportunities and risks of social peer review – as a special case of online rating systems –, here, we recall different forms of evaluating scholarly publications. Since peer review is the pervasive instrument in the quality assurance of scholarly publications, most researchers will have their own experiences with the process and will probably have noticed some of its merits as well as some of its downsides. After their publication, articles can be evaluated through their usage – for instance, citations or altmetrics that are interpreted as an implicit acknowledgment of impact – and through social peer review, which we focus on in this chapter. In the following, we provide an overview on four practices of publication quality assessment: the classic (closed) peer review, open peer review, implicit ratings, and social peer review.

### 9.3.1 Classic Peer Review

The classic peer review process is a form of explicit evaluation and typically takes place prior to the publication of a research paper. Manuscripts are submitted to a publisher, where editors assign reviewers (experts in the respective area) to it. The exact workflow varies between venues, for instance, in the degree of anonymity of reviewers and authors: In blind review, the authors do not know the reviewers' identities; in double-blind review, additionally, the reviewers do not know who the authors are; and in (the rare) triple-blind review, not even the editors know the authors' identities. Other differences include the way of deciding on a final verdict based on the individual reviews and the process after the (non-reject) decision, such as a rebuttal phase, several iterations of revising the manuscript, or simply a plain accept and publication. Many publication

venues use web-based management systems, like EasyChair,<sup>16</sup> for the reviewing process. Authors can submit manuscripts there, editors can distribute submissions to reviewers and collect the reviews to reach a decision. Often, referees can express preferences for submissions they would particularly like to review, or they can describe their expertise using keywords, which allow the system to match reviewers to fitting manuscripts. Upon submission of the reviews, referees can usually estimate the confidence they have in their own expertise regarding the specific manuscript.

#### 9.3.2 Open Peer Review

This variant of the peer review process is, like classic peer review, a pre-publication process. The difference to the above is that the reviewing process is made public. The openness can have different forms: One option is to open the circle of reviewers, allowing the submission of reviews without invitation by an editor. The reviews can be made publicly available, for instance, on the publisher's website or as appendix to the articles. The editors can keep the reviewers' names secret, or disclose them, or they can leave the choice to the respective reviewers. Furthermore, open peer review platforms can feature means to respond to the criticism and to engage in discussion with the peers.

#### 9.3.3 Implicit Ratings

Implicit ratings are a form of post-publication evaluation. The extent of a publication's usage is measured in terms of citations or with altmetrics, like counts of downloads, postings, publication-related tweets, and so on. In contrast to the two previous evaluation forms, here, actually rating a publication is usually not the primary purpose. Instead, using a publication is interpreted as evidence for its quality or impact, making it an *implicit rating*. The resulting metrics range from simple counts to more complex measures, for example measures that consider the source of the implicit rating (e.g., the impact of a citing publication as indicator for the relevance of the cited publication). Several measures, like the Journal Impact Factor (see Section 2.2.1) or an author's h-index [Hirsch, 2005], are based on citation counts, where receiving many citations is understood as evidence for high impact. The situation is similar with the web-based altmetrics. A selection of such measures in scholarly tagging systems has been proposed and discussed by Taraborelli [2008]. In this thesis, we have investigated several metrics for BibSonomy in Chapter 6, where we also visit literature on altmetrics in other publication management systems. A number of further data sources for implicit ratings is discussed by Priem and Hemminger [2010] and by Thelwall [2012], among them blogs and microblogging systems, recommender systems, citations in Wikipedia, or comments on articles. A critical drawback of the implicit ratings approach is the interpretation aspect. While at first glance it seems plausible that a heavily used publication must be relevant, individual usage events need not necessarily be

---

<sup>16</sup><http://www.easychair.org/>

attributable to the publication's quality. A publication can be cited when it is refuted or at least argued against; downloading an article does not necessarily imply that it is even read. Reacting to that and other issues, the Leiden Manifesto [Hicks et al., 2015] lists ten principles for research on usage metrics, the first of them being that quantitative evaluation should be considered rather as additional support for qualitative evaluation by experts.

### 9.3.4 Social Peer Review

Social peer review is a post-publication process where literature that has already been published, is subjected to the critical opinion of readers. In a social peer-reviewing system, readers can write a review or rate the publication or particular traits (e.g., its interestingness, novelty, or technical validity). The social aspect comes from the idea that reviews are not only provided by invited referees, but by the system's users on their own accord. Thus, social peer review is not a tool to decide whether or not an article is published. However, the results can be highly relevant for a publication and its authors, for example, when critical issues are discovered that were missed during the (pre-publication) peer review. Furthermore, these explicit (as well as the implicit) ratings can be used to rank publications in search engines and, thus, they can increase or decrease a publication's visibility. While the opinions or the resulting aggregated ratings have no influence on whether or not an article is published, they very well can influence whether or not it is read (and subsequently cited). Particularly high (low) ratings can attract (repel) users. Moreover, like the citation or altmetrics scores, the ratings can be used to rank publications in search engines and recommendations. Finally, the discovery of critical issues that were missed during the (pre-publication) peer review, can even lead to the publication's eventual retraction. For instance, Mandavilli [2011] reported a case where a publication met serious doubts from researchers on Twitter and the publishing journal subsequently issued an "Expression of Concern" (meanwhile, the paper in question has been retracted).

Aside from assessing a publication's quality, social peer-reviewing systems provide the opportunity to simply comment on a publication, for example, to add related work, to list errata, or to start a public discussion with the authors. Social peer review can also be integrated as a secondary feature into publication-related web systems, for instance, into scholarly tagging systems, like BibSonomy or CiteULike. Finally, there are systems, like Facebook or Twitter, where social peer review is conducted more informally, for instance, in a tweet with only a short comment. In these cases, there is no explicit structure for reviews and no explicit ratings that could be aggregated into an overall quality score.

## 9.4 Opportunities and Risks of Social Peer-Reviewing Systems

The purpose of peer review is the quality control of scholarly publications. Its goal is to avoid erroneous, sloppy, or irrelevant work and thus an abundance of unnecessary publications. Furthermore, it provides the opportunity to point authors of (already) good manuscripts to possible improvements before publication. However, particularly against the background of an ever-growing flood of submitted manuscripts, the ability to ensure publication quality has been questioned. Social peer review, as a post-publication quality assessment, can at least mitigate some of peer review's drawbacks, and therefore, it is the central topic in this section.

In traditional peer review, the peers are experts who are selected based on their reputation in a particular research discipline. For the more open social peer review it is not per se clear who the peers should be. We discuss different choices in the first part of this section. In the second part, we comment on opportunities and risks in the technical design of social peer review portals and compare to similar aspects in the other three forms of publication quality assessment (those mentioned in the previous section).

### 9.4.1 Opportunities and Risks of Choosing the Peers in Social Peer Review

The evaluation of research requires researchers to conduct it. At first glance, making quality control a research-internal process seems an intransparent approach. It would be preferable to have an independent third party as judge. However, it becomes clear that, due to the nature of research, only other researchers are actually qualified. One goal of the selection of reviewers in the classic pre-publication peer review, where manuscripts are accepted for publication based on the reviews, is to ensure that the referees possess the necessary qualifications to judge the submitted manuscripts. Hence, experts in the respective fields are asked to provide the reviews. They are the manuscript's authors' peers in the sense that they are experts in the same domain. The experts are chosen and invited by the publication venue's editors, who know them at least by reputation. In an open online social peer-reviewing system, usually, no authorities like editors exist. Thus for such systems, there must be another way to select peers. Attending to Research Question RQ2, we discuss possible choices for selecting peers.

Already by name, a characteristic of social peer review is the combination of *social* and *peer*. The term *social* stands for the interactive, collaborative processes between a system's users that characterize the Web 2.0, like in social networks, or social tagging systems. In these portals, users are usually welcome as long as they abide by the terms of use. For a *peer* review, however, additionally some level of expertise should be required from the users. Following the arguments in [Doerfel et al., 2013c], peers should be holders of the freedom of sciences (Section 2.5), thus, persons who actively

engage in research activities on their own responsibility or plan to do so. This includes particularly researchers (employees of universities or other research facilities) but also covers students who conduct research (e.g., in their bachelor or master theses).

Editors of journals, where the publication of an article implies a certain prestige for its authors, will choose a much stronger restriction to determine peers (e.g., well-known colleagues from the respective area of research). However, in the context of social (post publication) peer review, the restriction mentioned above seems more appropriate. It allows interdisciplinary discourse and stronger restrictions would hinder the social aspect of the process. In the following, we exemplarily present several ways to implement an assessment of whether or not a user of a social peer-reviewing system is a peer and thus suitable to review a publication. For each option, we discuss opportunities and risks for the involved parties.

The first question the operator of an online peer-reviewing system will have to answer, is how to recruit the raters. Following the example of traditional (pre-publication) peer review, referees could be invited. Different modes for this option include invitations issued by a publication's authors, by the system's operators (who would thus assume the role of editors), or using some trust system, such as allowing already invited reviewers to invite further reviewers. Considering the immense efforts of such undertaking, it seems reasonable to consider an alternative method: an open peer review where the users themselves choose the articles they wish to comment on. Of course, a reviewing system could also support both modes for acquiring reviewers. The following access management features mainly pertain the open peer-review scenario. Here, we distinguish between *overall* and *publication-specific* qualifications. The former means the qualification to provide reviews in general, while the latter means the expertise to judge a particular publication.

### **Voluntary Report of Own Qualifications**

By a voluntary report of one's own qualifications, we mean that users enter data which can be used to judge their expertise. Users could be required to make an overall assessment of their qualifications or to make publication-specific statements. For an overall report, users could declare their qualification during the registration, for instance, their academic degree (chosen from a pre-defined selection) or their affiliation with some research institute. Furthermore, users could be required to select those research disciplines in which they are qualified to provide reviews. The system can then restrict the submission of reviews to publications from these areas. To ensure that the qualification assessment is up to date, the rating system could remind users in regular intervals to confirm or update their qualification profile in the settings.

A publication-specific version of this option is to ask reviewers about their qualifications before they submit a review for a specific publication. This variant is close to the classic peer-review scenario where reviewers usually have to estimate their confidence in their own expertise regarding a publication's topics. This form is more fine-grained than the overall approach, however, it also requires more effort from the users.



Both variants allow for defining the group of peers. A simple discrimination could be based on a user's academic degree and fields of research. Furthermore, the data can be used to weight individual judgments in an aggregation of ratings. For instance, reviews from professors might have a higher weight than those of students. In the publication-specific version, ratings could be weighted with the confidence scores of reviewers.

While the aforementioned processes seem technically easy, there are some details that pose serious obstacles: Academic degrees are not standardized. Many different degrees and titles can be listed in a corresponding selection, but it is hardly possible to achieve completeness. In consequence, users will have to select the most similar degree to their actual one. In addition, comparability of degrees issued in different countries is non-trivial. An obvious problem with self-assessment is that fraud is easily possible. In the classic peer-review process, providing false information would pose a serious threat to one's own reputation (if discovered). Similar social pressure can, however, not be expected in a social peer-reviewing system, particularly when the reviews are submitted anonymously. A second related problem is the subjective nature of self-assessment, resulting in a (possibly subconscious) self-reporting bias. Whether or not users tend to under- or overestimate themselves can hardly be determined.

### **Certified Qualification**

To avoid the aforementioned problem of easy fraud, a system could require its users to actually certify their qualifications. For example, users could be asked to provide a copy of their degree certificate. Although this method would be relatively safe – fraud would require the document forgery – it is not suitable for an open social peer-reviewing system. Since there is no general standard for such documents, all registrations would have to be checked manually by an expert. Such effort, as well as the implied privacy issues, do not seem to be justified by such a system's purpose.

### **External Sources**

Another variant of the verification – or rather estimation – of a user's qualification is the utilization of external sources. Digital catalogs, like DBLP,<sup>17</sup> or scholarly web search engines, like Google Scholar or Microsoft Academic Search, allow querying for scholarly publications by an author names. They could be used to automatically generate user profiles, for example, including previous or recent publications, as well as citation and publication statistics. Such profiles paint a picture of a user's expertise in general, and they could also be matched against individual publications to determine a publication-specific level of expertise. Moreover, a system could interpret the received statistics as estimators for the expertise of a user and use these values not only to determine whether or not a user is a peer, but also to weight reviews according to the estimated level of expertise.

---

<sup>17</sup><http://dblp.uni-trier.de/>

For this procedure to work, users would have to reveal their real names truthfully to the system, opening the possibility of fraud. Furthermore, the acquired statistics are hardly comparable in different research disciplines with different publication cultures. Other problems with this solution arise from the inclusion of third-party data. Publication collections, like the ones mentioned above, usually contain only a subset of the actually existing literature, resulting from the mode of acquiring the data (e.g., crawling web sites versus systematically including publisher data). Student's theses are rarely included, publication data can be erroneous or incomplete or suffer from the automatic extraction from their original sources. Moreover, authors can change their names (e.g., in marriage), and authors can have identical names, making it difficult to distinguish them. Finally, collecting data from external sources requires particular care regarding privacy protection (see Doerfel et al. [2013c] for details).

### Deriving Expertise from User Activities

The last variant for determining user expertise we discuss here, relies on system-inherent data: the data users create through their activities in the system. This includes previous reviewing activities but also other interactions with the system. For instance, in a combined tagging and reviewing system, like BibSonomy, users have their own collections of tagged publications (including those they authored themselves<sup>18</sup>), their network of friends, their clicks, downloads, search queries, and so on. From such information, one can extract a user profile describing a user's interests and expertise. Furthermore, other users can contribute to a user's profile. By rating reviews (similarly to the "helpful" ratings known from Amazon (see Section 9.2.3) or – in tagging systems – by posting publications that have been written by other users.

While such a system does not require the users to provide explicit information on their own expertise, it still relies on information provided by the users, albeit implicitly. Thus, it opens the door for fraud by spamming. For example, if "helpful" ratings were counted as indicators for a user's expertise, others could use them to increase or decrease a particular user's reputation. Similarly, in a scenario where the tagged resources are used to determine expertise, users could easily improve their profile by selectively adding resources to their collections. Here, it will be the system's operators' challenge to implement appropriate spam detection techniques and safeguards against the creation of multiple accounts. From the judicial point of view, the here described profiling of users bears the risk of affecting the users' right to informational self-determination.

#### 9.4.2 Opportunities and Risks of Social Peer Review

Peer review is the established quality control in scholarly literature. However, there are aspects that have been criticized, and even its abolition has been proposed. For instance, Mandavilli [2011] cited Cameron Neylong (researcher and author of the

---

<sup>18</sup>BibSonomy encourages its users to tag such publications with the tag "myown".

blog “Science in the Open”<sup>19</sup> with the statement: “It makes much more sense in fact to publish everything and filter after the fact.” In literature, various authors have discussed arguments in favor and against traditional peer review and its open alternatives [Weller, 2004, Hornbostel and Simon, 2006, Müller, 2008]. Since our focus here lies on the opportunities and risks of social peer review, we only briefly list some of the aspects that have been criticized in the classic peer review model and that have to be examined for social peer review as well. For further details, we refer to the above mentioned sources. Many of the aspects that we discuss in the following, are widely known and they rely on experiences that probably many researchers had with their own manuscripts and publications. Therefore, the contribution in this section is less the enumeration of these issues, but rather the juxtaposition with social peer review.

Following the example of Müller [2008], we consider various aspects for discussion, roughly adapting and extending the set of categories there. We compare the four forms of assessing scholarly publications – closed and open peer review, implicit ratings, and social peer review (see Section 9.3) – particularly focusing on social peer review. We thus address Research Question RQ3 and discuss the opportunities and the risks that each form presents regarding the aspects of time, motivation for reviewers, learnability of reviewing, cost, fairness, transparency, and manipulation possibilities.

#### **Aspect: Time**

A frequently voiced criticism of the classical peer review is the long time it takes from a manuscript’s submission to its publication. While for conference submissions the conference scheduling sets deadlines for the reviewing process, journals often do not have fixed deadlines, and authors have to wait several months before they are informed of a decision. When revised versions are requested, these have to be prepared and resubmitted by the authors and are then again reviewed. When a manuscript is rejected, the authors can submit to a different venue and thus restart the process.

With open peer review, the time issue is similar. However, if reviews are made publicly available after the reviewers have submitted them, authors have the opportunity to react earlier to the criticism and to discuss or to adapt their work accordingly. Moreover, if reviews can be submitted by others than just the invited experts, their opinions can also shorten the time of waiting for the authors.

Implicit ratings occur after publication and thus do not influence the time it takes to get a manuscript published. Rather, it takes time for the measured usage actions to accumulate. While in altmetrics responses can occur early after the publication, citations take much longer since citing publications have to be written and of course peer reviewed themselves. For example, the (journal-level metric) journal impact factor (cf. Section 2.2.1) considers citations within two years after publication.

---

<sup>19</sup><http://cameronneylon.net/about/>

**Social Peer Review.** In contrast to pre-publication peer review, social peer review is not constrained to a particular time-frame. This is especially an opportunity for older publications or articles from less prominent venues. When and if a paper receives comments, depends (among other things) on how much the research community is interested in its topics. Thus, long forgotten articles can come into the focus of current research again. On the other hand, reactions can occur extremely fast. Mandavilli [2011] described a case where an article was discussed on Twitter, leading the authors to concede technical errors within one week after publication. However, she also noted that the large majority of papers is never discussed at all.

#### **Aspect: Reviewer Motivation**

By providing reviews, researchers have the opportunity to actively shape their respective fields by bringing forward, by improving (through constructive criticism), or by hindering the publication of the manuscripts they judge. When referees are invited to review a manuscript, they are usually not payed in money, but in reputation, which is gained from being a member of a journal's editorial board or part of a conference's program committee. This "reward" does usually not (at least not directly) depend on the quality of the work, and it is granted before the reviewing process even begins.

An explicit problem for the motivation can be the large number of reviews a referee is asked to provide. This is particularly the case when a researcher works for several venues or when many submitted manuscripts must be distributed among only few reviewers.

In open peer review, the motivation is similar for the invited reviewers. For others, the motivation can be to profit from the early availability of new results and the opportunity to influence them before they are published, for instance, by pointing out errors that can be fixed before publication. The publication venue for which the reviews are provided may choose to create additional incentives, such as naming them as experts and contributors or giving away awards for particularly distinguished referees. For instance, the Semantic Web Journal honor reviewers of their open peer review process annually.<sup>20</sup>

For implicit metrics, the motivation question does not arise, because the users act in their own interest and the rating is only an interpretation of their actions.

**Social Peer Review.** In contrast to the classic peer review with invited referees, in social peer review, reviewers act on their own impulse. Motivations for contributing include pointing colleagues to particularly valuable results, warning about wrong conclusions or technical errors, hindering follow-up research building on erroneous foundations, pointing to own research, and asking questions with respect to a publication. Moreover, the review portal can add additional incentives, like rankings of the top reviewers, based on their contributions or on comments responding to their

---

<sup>20</sup><http://www.semantic-web-journal.net/blog/semantic-web-journal-awards-2015> (accessed October 21, 2015)

reviews. In that manner, researchers could receive similarly visible appreciation as with memberships in program committees or editorial boards. Such incentives are in use, for example, in the web shop Amazon.<sup>21</sup>

### **Aspect: Learnability of Reviewing**

The ability to write scholarly reviews is often acquired “learning by doing”, for example, as a sub-reviewer for someone else, who can serve as a mentor. Moreover, as reviewers are peers, they are authors as well. The reviews they received for their own manuscripts are experiences they can tap into when writing reviews themselves. Only rarely, referees receive feedback for their reviews. An exception are procedures where the authors can respond to the reviewers. A source to compare one’s own reviews to, are reviews from colleagues on the same manuscript. The more open the review process is arranged, the easier it is to find examples for good reviews as advice. This is one strong advantage of open reviewing, particularly for beginners. For implicit ratings, the aspect does not apply.

**Social Peer Review.** Like open peer review, this form of reviewing offers the opportunity to study other users’ reviews and to find both good and bad examples. These can be identified by the portal, for instance, through helpful-ratings from users or by extracting information about their quality from comments responding to them.

### **Aspect: Costs**

Another significant point of criticism are the costs of peer reviewing, both open and closed. They originate at the publication venue. However, it must be taken into account that the reviewers are working free of charge, and that the writing of reviews is part of their activities on the job. Thus, the time invested must be considered as hidden costs that are paid by the referees’ employers (i.e., the research institutions they work for). The costs are especially high when the manuscripts are only partly related to a researcher’s own work and more time is required to familiarize with the field. Moreover, a scaling problem has arisen from the growing numbers of publications [Bornmann and Mutz, 2015] and the accordingly increasing number of submitted manuscripts. Taraborelli [2008] mentioned the issue of scaling as one of the major drawbacks of peer reviewing: Through the strong increase in manuscripts, researchers have to devote more and more of their time to writing reviews.

A second cost-related problem arises from the opportunity to submit manuscripts that have been rejected to other venues. One strategy to deal with negative reviews is simply to submit the manuscript to another venue (often of slightly lower reputation). This can be repeated until finally some venue will publish the manuscript. While the additional reviews should be used to improve the manuscript, they still must be counted into the overall cost of the peer-reviewing process, which then are far

---

<sup>21</sup><http://www.amazon.com/review/top-reviewers>

higher than just the time invested by those reviewers who finally voted to accept the publication. Moreover, the newly assigned reviewers cannot profit from the work of their colleagues who previously already pointed out critical issues in a manuscript.

Costs of implicitly generated evaluations arise mainly from the computational effort and from collecting the data. Such datasets must be sufficiently large to yield reliable, representative statistics. For example, for citation-based metrics, the records must either be collected manually or using automatic techniques, like crawling and extracting publications from publisher websites or the researchers' own homepages. Depending on the quality of the data, cleaning and correcting it can be another expensive process.

**Social Peer Review.** Costs accrue for the system operators who run the reviewing portal, finance the hardware, software, maintenance, and provide service to the users, like answering questions or reacting to complaints. Further costs incur for the referees in the form of time that they invest into writing reviews. In contrast to classic peer review, however, they do not review papers assigned to them but rather literature that they encountered during their own research and which they wanted to read anyway. This lowers the additional effort of familiarizing oneself with a papers topics as the reviewer is often already invested in the respective area and does not have to judge papers that only peripherally touch their own field of research. Finally, it is up to the reviewer to write a full-blown review or simply to leave a brief comment.

#### **Aspect: Fairness and Validity**

A particularly important criterion is that of fairness and validity. A valid procedure is expected to yield reliable results, meaning that, at least to a large extent, reviewers will agree in their opinion and that applying the peer-reviewing process twice will yield a similar outcome. However, unreliability is one of the aspects that often is criticized – Müller [2008, Section 4.2.3] listed several critical studies. For authors, an unfair reject is highly problematic, as it takes time to resubmit the manuscript to some other venue where the reviewing process begins anew, hindering follow-up research.

A factor that promotes validity is that the editors can choose referees according to their qualification with respect to the manuscript. When the chosen referees decide to hand the task to someone else – a subreviewer – it is their responsibility to ensure the review's quality. Still, Campanario [1996] listed several examples of highly successful articles (measured in terms of citations) that had problems during the publication process and that have been rejected at various venues before they were finally published. In contrast, the blog Retraction Watch<sup>22</sup> repeatedly reports cases of articles that passed peer review and were published but have been retracted by the publishers later.

A consequence of choosing peers as reviewers is that often authors and referees belong to the same research community and therefore are likely to be colleagues (or competitors). This can make it difficult to stay impartial. Especially judgments

---

<sup>22</sup><http://retractionwatch.com/>

in categories like “interestingness”, “relevance”, or “expected impact” – which are typically part of a review assignment – are not fully objective and likely to reflect the reviewer’s personal opinion. The choice of the reviewers can produce further biases (towards nationalities, gender, established researchers, etc.). An attempt to reduce biases towards favorable reviews for colleagues is the option for referees to declare conflicts of interest with colleagues or institutions and, thus, to avoid having to review close colleagues or collaborators.

An interesting situation arises in open peer review, when referees sign their name to their reviews. Without anonymity, reviewers can be biased towards more sympathetic judgments. On the one hand they might fear receiving negative reviews for their own work, as a form of retaliation, and they might generally expect resentments from colleagues and strained collaborations. For both authors and referees, open peer review bears the risk of being openly criticized for own mistakes: Referees must point them out as part of the review and on the other hand, authors might respond to erroneous reviews.

Greaves et al. [2006] reported an experiment of Nature in 2006 where submissions were subjected to traditional (closed) peer review, as well as to an open online peer review. Only few researchers actively contributed to the process and the editors, judging the value of the comments, felt that they were rarely helpful. However, open peer review is successfully conducted elsewhere, like at the Semantic Web Journal, where reviews are usually published non-anonymously during the publication process.<sup>23</sup>

The fairness of implicit ratings depends on their technical setup, including the metric, the data source, and the extraction processes. We already saw in Chapter 6 that different measures – even on the same dataset from the same system – are not necessarily correlated with each other, nor with impact measures, like citations. A positive aspect of implicit ratings is that the evaluation does not rely on only few experts but on a much larger group of people. For instance, citations can be contributed by all authors of scholarly literature, and altmetrics measures can be influenced by basically anyone with access to the web and, thus, also reflect impact beyond the research community. At the same time, the broad crowd of possible “referees” is also a problem, as the group of users can hardly be restricted or controlled. Those who compute measures from publication usage, must ponder between completeness – including as many usage events as possible – and a restriction on the set of admissible users or publications: For example, citations counted on the Web of Science come from manually selected publication venues, citations on the web search engine Google Scholar are collected from scholarly publications found on the web, and altmetric statistics count downloads, postings, or mentions in open web systems, like Twitter or Mendeley. Moreover, user-generated content – citing publications as well as resources on the web (e.g., publication posts in a tagging system) – suffers from other issues, like unclean data (typos, missing entries, etc.). Another drawback of using implicit ratings is that they do not yield productive feedback in terms of actual criticism or suggestions

---

<sup>23</sup><http://semantic-web-journal.net/reviewers> (accessed October 23, 2015)

for improvements. A number of drawbacks of bibliometric measures is discussed by Glänzel and Moed [2002], among them statistical issues, differences between various research areas and the choice of parameters in the metrics. The same arguments can essentially be made for web based metrics as well.

**Social Peer Review.** The degree of fairness depends on the degree of anonymity, much like in pre-publication peer review. A difference arises from the wider circle of potential reviewers. While in classic peer review, peers are chosen by the editors, the group of peers who can write reviews in a social peer-reviewing system is much larger (see Section 9.4.1). An advantage could be that more reviewers provide a more representative judgment about a publication. However, as mentioned above, it is difficult to distinguish the peers among all potential users of an online rating portal. Allowing non-experts to judge publications is a mixed blessing: On the one hand, it seems unfair to allow unqualified readers to publicly criticize the research. On the other hand, it can give a voice to practitioners who profit from the research: A new method can be applied by others without necessarily understanding the mathematical proof of its validity; results of studies can be used without necessarily understanding the exact experimental setup. If information about the raters is available (e.g., volunteered information about educational achievements, fields of expertise, etc.), it can be used to filter ratings and to produce different aggregations of the ratings, for instance, aggregating only expert ratings versus all ratings a publication has received.

A second aspect that is different from classic peer review is that the referees can come from various research disciplines (in pre-publication review, they are typically chosen from the respective fields, determined by the manuscript). This entails that referees from other fields may apply other standards than referees from the same area of research would. On the other hand, the possibility of interdisciplinary discourse is a welcome aspect of social peer review. Moreover, similar to filtering by expertise in general (mentioned above), reviews can also be filtered by the research areas of the reviewers, for example, displaying only reviews by experts from the discipline of the publication. By enforcing a fix structure for reviews, for instance, subcategories such as technical validity, presentation, impact, and so on, it would also be possible to compute aggregated ratings only per category and to allow authors to leave some categories unrated. Thus, authors from outside a publication's research area could value the applicability of a result in their field, but would not have to make a judgment on the validity of the results, which can be hard to judge for outsiders.

A risk of social peer review poses the possibility that despite the large number of potential reviewers, without explicit invitation, only few publications are reviewed and individual publications only receive few ratings. Thus, individual opinions can have a high impact on a publication's reputation in the system. Furthermore, aggregated ratings – which could be used to rank publications (e.g., in search results) – are less meaningful when they rely only on few contributions. For example, using the arithmetic mean to aggregate numeric ratings, publications with one very good review



would be ranked higher than publications with many very good reviews and one that is “only” good. In statistics, other aggregations are known that include the distribution of ratings, the previous ratings of the raters, as well as confidences (for the aggregated rating being an estimate for the theoretical result, if all users had rated). Several models were surveyed in [Jøsang et al., 2007, Section 8]. Techniques for the filtering of particular unfair reviews have also been proposed (e.g., Whitby et al. [2004]).

A drawback of publicly available results (in social peer review as well as in open pre-publication peer review) is that after the first review has been published, others will be influenced by its statements. Thus, reviewers might already be biased by the opinions of their predecessors. The drawback regarding fairness is, however, an advantage regarding time, as it can save both authors and reviewers time if the same critical aspects are not repeated due to the unawareness of other reviews.

### **Aspect: Transparency**

In the traditional peer-reviewing process, transparency depends mostly on the level of anonymity, which protects referees and, in double-blind settings, also authors. Despite the omission of names, it can be possible for reviewers to identify the authors, for instance, by analyzing the cited authors, by recognizing a particular style of writing, or simply from familiarity with the key players in a particular discipline. Vice versa, for authors it is more difficult to guess the reviewers’ identities. However, even in (single-)blind review, where reviewers know the authors’ identities, the process is not completely transparent from the referee’s point of view: They do not know if a manuscript has previously been submitted elsewhere and why it was rejected earlier. Thus, the effort of the previous reviewers is lost and the new referees have to start from the beginning.

Transparency is one of the strong suits of open peer review. If the reviewers publish their evaluations together with their names, the only thing that remains unknown are the editor’s reasons for choosing the referees they invited. Since authors know their reviewers, they can even compare the reviews they received to other reviews by the same referee.

In implicit evaluations, metrics accumulate usage data and it depends on the provider of these metrics to what extent their process is transparent. The span ranges from black box, like on ResearchGate<sup>24</sup> [Kraker and Lex, 2015], to more transparent measures like Google Scholar’s counting of the citations they encounter in web crawls. The latter method allows users to verify the resulting scores by following the linked citing publications. Depending on the data that the metrics are based on, such validity checks can, however, easily become rather time-consuming. More complex metrics do more than just counting usage events (citations, downloads, etc.) and incorporate other influences, such as the age of an article, the citations of a citing article (the impact of the influenced research), or the position in the citation network (e.g., *PageRank*-like

---

<sup>24</sup><http://www.researchgate.net/>

methods). With such metrics it is no longer possible for individual users to verify them. However, even with simple metrics, the collection of the data is a source of intransparency. For a user, it is unclear why particular sources are chosen to gather the data, for example, which websites are crawled to collect publications, or why particular web systems are chosen to count usage events. Finally, depending on the chosen metric, the matter of who “assigns the ratings” (i.e., who are the users), can be a source of intransparency. While for citation-based scores, it is known who the citers are (the authors of the citing publications), altmetric approaches do not necessarily have that level of transparency. For instance, for counted mentions in social networks, it is unclear who the users are due to the use of nicknames. The Leiden Manifesto [Hicks et al., 2015], which lists ten principles for the design and interpretation of usage metrics, focuses on issues of transparency particularly with the rules “Keep data collection and analytical processes open, transparent and simple” and “Allow those evaluated to verify data and analysis.”

**Social Peer Review.** The level of transparency in social peer review depends on the degree of anonymity for the referees. The authors cannot be anonymous (like in double blind peer review), since their names are obviously connected to the publication that is reviewed. The highest degree of anonymity would be achieved when users can add ratings without signing on to the system. However, users usually are required to register with some chosen pseudonym, the user name, and it is up to the users to make the association between real name and user name easy or difficult for others. The level of transparency can be raised when the system requires its users to provide their real name to the system, or even to prove their identity. Transparency is already higher, when the real names are known at least to the system (and not necessarily to the users) since thus, multiple registrations can be avoided and it becomes easier to assess the reviewer’s status as an expert or peer (see Section 9.4.1). Furthermore, it allows for contacting reviewers in cases of alleged misuse or libel. The advantages and drawbacks of anonymity are similar to those in pre-publication peer review, particularly the risk that referees might be unwilling to utter criticism when they are not protected by anonymity.

Transparency is not only a matter of reviewer anonymity. Facilities for publishing criticism, for answering to such criticism, or for discussion in general, enable a transparent scientific discourse. Errata, supplements, or references to related work can be added to publications and are, henceforth, available to future readers. While in classic (closed) pre-publication review, only the authors can profit from the reviews, in social peer review, reviews become available to a larger public on a long-term basis. An idea that advances the classic way of handling publications even further (in the same spirit) is that of *liquid publications*, where publications are no longer fix articles, but evolving “science knowledge objects” [Casati et al., 2007] that can be updated and improved much like open source software.

For readers, authors, and reviewers, a problem arises when several systems are used for social peer review. While in classic peer reviews authors know that and when their work is reviewed (after all, they submit the manuscripts to be peer reviewed), in social peer review this is no longer the case. An author's work can be publicly criticized without the author's knowledge. This makes it difficult to react. To compensate, systems could send alerts when publications receive new reviews, for example, using the authors' email addresses that usually can be found in the publications. The latter is, of course, only possible, when the full texts are available in the system. Other researchers could profit from such services as well, for instance, by being alerted when publications that they find interesting receive new comments.

### **Aspect: Manipulations and Misuse**

Publications that have passed the (pre-publication) peer review are important stepping stones for the reputation of their authors. Intransparent processes, however, provide opportunities for gaming them for one's own purposes. When reviews are not disclosed, it is easier to be overly harsh (unfair competition) or forthcoming (nepotism). In open peer review, it becomes easier to identify such misuse and non-invited reviewers can add their opinions, countering potentially unfair reviews.

Opportunities to game metrics arise in all implicit-rating scenarios. Self-citations, or citation cartels (colleagues who frequently cite each other) help raise citation counts. Falagas and Alexiou [2008] exposed ten ways for editors to improve their journal impact factor. From these follow ideas for authors to make their manuscripts more attractive to editors. Ferguson et al. [2014] described cases of explicitly cheating peer review, such as tricking editors into assigning authors as referees for their own manuscript. A practical experiment by López-Cózar et al. [2012] showed how to easily boost author-level citation counts on Google Scholar, using fake publications. Finally, usage-based metrics can be gamed using multiple accounts or scripted, simulated usage (e.g., producing downloads of publications or page visits). Even explicit gaming aside, implicit evaluations through usage metrics are an incentive for researchers to optimize towards them (e.g., in selecting journals, placing citations, using particular tools, etc.) and, thus, an influence on researchers and their work.

**Social Peer Review.** To be accessible worldwide, a social peer-reviewing system must be web-based. Thus, it is vulnerable to the same well-known attacks and bears the same risks as any web system. The potential of manipulated ratings and of libel or inappropriate criticism exists, as described in Section 9.2 for arbitrary online rating systems.

With higher transparency, especially when reviewers have to reveal their real name or at least their affiliation to the system, the risk of manipulation can be reduced and discovered more easily, and the creation of multiple accounts by the same person is significantly more difficult. In contrast to the established pre-publication peer review, where discovered manipulation can have severe consequences (retracted publications,

damaged reputation), the situation for social peer review is less dramatic. However, it is incumbent on the system's operators to punish manipulations. For details on liability for disturbance see [Doerfel et al., 2013b, Chapter 8].

Next to technical means of manipulation, also the other phenomena, like nepotism, which are problematic in peer-review, can occur in social peer review as well. The more open the reviewing system is, the easier such forms of manipulation can be detected. The system could, for example, check previous collaborations (co-authorships, extracted from the publications in the system) to identify connections between raters and the authors, and red-flag suspicious cases. Furthermore, the system could invite authors to notify its operators if they discovered (alleged) fraud. Allowing the rating of reviews (helpful ratings) would be another way to counter unfair reviewers. However, even that procedure can become a target for manipulation, as observed on Amazon,<sup>25</sup> where reviewers deliberately tried to damage the reputation of other reviewers to outrank them on a list of the top reviewers.

## 9.5 Realizing a Social Peer-Reviewing System

In the previous section, we have discussed various features with regard to opportunities and risks. In this section, we address the question of who should run such a system. Having it operated by the research community or by some dedicated institution with high reputation, would help the system's credibility and provide the necessary sincerity. In the first part of this section, we discuss this prospect. A way to increase the interest in such a system could be to integrate it with tools that researchers use anyway for their work with literature. The second part of this section, therefore, addresses implementation issues of social peer review in a tagging system for scholarly literature.

### 9.5.1 A Social Peer-Reviewing System Operated by the Research Community

The previous discussion shows that none of the four analyzed forms of publication evaluation are without problems in all considered aspects. However, all these forms have different advantages, and they can complement each other. Therefore, it seems reasonable to use all of them in a two-stage process: Manuscripts are subjected to open or closed peer review before they are published; afterwards, implicit ratings as well as social peer review are employed to further assess the publications' quality. Implicit ratings, citation-based or altmetrics, are integrated in various ways into the established systems. Often, the data on which computations are based, is property of companies, like Google or Thomson Reuters (Web of Science), and it is up to them to decide which publications to include. This decision is either made explicitly, like in the case of the Web of Science, where editors decide on a set of journals (and more recently

---

<sup>25</sup><https://web.archive.org/web/20130410163258/http://www.readers-edition.de/2009/10/01/amazon-schmutzige-klicks-gegen-top-10-rezensenten/> (accessed October 28, 2015)

also conferences), or it results implicitly from other design decisions, like in the case of Google Scholar, where the included documents are determined by the chosen crawling strategy. Metrics like downloads, mentions, or posts, are by nature constrained to the system they occur in, (e.g., the digital library or the social bookmarking system). Aggregators can compensate by collecting these altmetrics from various sources and presenting summaries (see [Jobmann et al., 2014] for a critical comparison of four popular altmetric providers).

Social peer-reviewing systems rely on explicit rating activities. An imminent danger is low participation. Discussions on scholarly publications do occur in social media, however, from there they are difficult to find (especially after some time has passed) and it can be difficult to match them to the publication they concern. This issue could be solved with a central system that is run by or on behalf of the research community itself. In such a system, all (or as many as possible) scholarly publications – peer-reviewed publications as well as previously unreviewed manuscripts or preprints – would have to be added and could then be reviewed. A committee could (partially automated) accept and add documents and their metadata. Reviews could be added on own initiative or upon request. For example, authors might want to invite colleagues who cited their work to write an opinion.

Reviewers could be granted different levels of anonymity. Particularly, the system could require users to choose a pseudonym but to provide and to validate their affiliation to some research institution or even their real names. Users could still be allowed to post reviews anonymously or under their chosen pseudonym, but they could also choose to add their real names to some of their reviews. By offering these different levels of anonymity it could be possible to lower the barrier for reviewers to submit to the system. However, since their real names are known to the system, it would be possible to avoid the registration of multiple accounts by the same person and to limit the circle of potential reviewers. The fact that the real names of researchers are connected to their reviews in the system (albeit not visibly for the users) still bears the danger of being discovered as author of a particular criticism. However, this situation is similar to classic pre-publication peer review, where the reviewers are unknown to the authors, yet the editors do know by whom each review was submitted.

The commitment of researchers would be awarded with reputation and respect, and possibly with further incentives for two reasons: to motivate writing new comments and to motivate using their real names. Possible incentives include public announcements of particularly active or especially helpful reviews, best reviewer awards, or even more complex reputation systems, for instance, following the example of question-and-answer systems like Stack Overflow,<sup>26</sup> where answers (reviews) can be voted up or down, etc. Ratings and reviews could be filtered and aggregated by different groupings: with regard to content, with regard to the reviewers' expertise and with regard to the reviewers' anonymity level.

---

<sup>26</sup><http://stackoverflow.com/help/reputation>

To attract researchers from all disciplines, the system could provide multiple user interfaces to choose from, offering different sets of features – depending on what is commonly used in different research disciplines. Furthermore, the system should provide public interfaces to access the data and to integrate with other systems. For instance, researcher who are active on Twitter, could be offered means to submit their tweets concerning a particular publication automatically as comment in the social peer-reviewing system as well; or, vice versa, to tweet a review (or rather a link to it) automatically when they submit it to the peer review platform. Similar tools can be created for other social networks analogously. Furthermore, the ratings and reviews could be shown in digital libraries, publication management systems, etc. Finally, if the system contains enough publications or at least their metadata including references, this corpus could be used to compute implicit-rating metrics, complementing the qualitative judgments of reviews with usage statistics.

Developing, maintaining, and operating such a systems would present a great effort. However, for several of the above described features, researchers have already constructed systems on the web, which can serve as models or building blocks for the proposed social peer-reviewing portal: A model for the editorially controlled collection of metadata into an online catalog is the system DBLP,<sup>27</sup> which is run at the University of Trier: Metadata is collected from the websites of publishers and users can advise the operators of missing journals or other venues. In the social tagging system BibSonomy, run at the Universities of Kassel, Würzburg, and Hanover, researchers collect references to those publications they use, write, or find interesting. The system also offers discussion and rating features that allow the submission of reviews and the assignment of numeric ratings (star ratings) to publications. Users can choose to review anonymously or reveal their user name. The visibility of reviews can be restricted to public (default), private, or to a particular user group. We discuss BibSonomy further in the next section.

The service ORCID is dedicated to provide unique identification of researchers – Open Researcher and Contributor ID (ORCID) –, solving the problem of name ambiguity. The service is maintained by the non-profit organization ORCID Inc., whose team is composed of scholarly publishers and members from various renowned research institutions. The ORCID would be one possible option for researchers to register with the system, preventing the easy creation of multiple accounts.

Collaborative reviewing systems of scholarly literature could provide a significant benefit, particularly against the background of the ever-faster growing corpus of scholarly literature. They can help systematize the available publications, increasing their online visibility, and filtering the most relevant content of the highest quality. Such systems can complement the pre-publication selection of peer review, but they are independent of it and can also be applied for literature that has been published at venues without peer review. Through feedback and commentary features, social

---

<sup>27</sup><http://dblp.uni-trier.de/>

peer-reviewing systems can support the scholarly discourse and provide qualitative justification for judgments of a publication's merit.

### 9.5.2 Realizing Social Peer Review in a Social Tagging System

A social peer-reviewing system is supposed to enable its users to submit and publish their opinions and to discuss and rate scholarly literature. In scholarly social tagging systems, like BibSonomy, Mendeley, or CiteULike, users already collect references to publications which they used or found noteworthy. Thus, they naturally suggest themselves to be augmented with reviewing and rating features. In the following, we discuss several issues with the implementation of social peer review in a tagging system, making BibSonomy our use case.

The content in tagging systems is user-generated, and each user creates their own posts. Thus, several posts (from different users) can exist for the same publication. To allow their aggregation, reviews and ratings must therefore be attached to the publications, not to the posts. Since publications are accessed through their posts, all posts for the same publication also belong to the same discussion (the set of all ratings and comments for one publication). While it seems only natural that users rate resources rather than posts, one must keep in mind that tagging systems are post-centric, meaning that almost all data belongs to a particular post rather than to a resource.<sup>28</sup>

In order to connect a discussion in a tagging system to a resource, there must be a way to decide which resources are the same, even though different metadata may have been entered in different posts. This problem is non-trivial: Users omit optional fields that others have added, use abbreviations for venues, or misspell or forget words. In BibSonomy, publications are identified by their title, year of publication, and their authors (or editors when no authors are present). These three fields are used to compute a hash that can serve as that publication's identifier. For details see [Hotho et al., 2006b]. Two publications are considered identical when they agree in these three fields. As a consequence, it sometimes happens that two similar but actually different papers are identified as the same. A typical example are extended versions of a paper (e.g., the journal version of a conference paper), since they often have the same title and authors as the original and, depending on the speed of the publication process, sometimes the same year. A drawback of this procedure is that it can happen that users review different versions of a paper and criticize aspects that have already been fixed in another version. A scholarly tagging system could also rely on external identifiers, like the digital object identifier (DOI). The use of the identifier would have to be enforced (i.e., users have to add it to their posts) and publications that do not have a DOI (yet) would be excluded.

Using the hashes as identifier yields another issue, arising from the possibility to update posts. This feature is necessary for correcting typos or for adding missing fields

---

<sup>28</sup>For example, users may post the same publication but each enter different metadata. Whenever users request that publication, they will see one of these posts and, thus, that post's metadata.

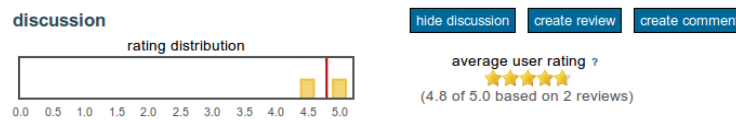
of metadata. However, when users change the title, authors, or year, the identifier of a publication will change. If the publication has already been reviewed, then either the discussion is attached to the new identifier of the post's publication or it stays with the old one. To compare the two options, consider the example that a post by user  $u$  containing the resource  $r_1$  is changed by  $u$ , who edits the resource. The result is a post containing the resource  $r_2$ . The option of keeping the discussion attached to  $r_1$  has two drawbacks: (i) The new post is no longer attached to any discussion, which is confusing for user  $u$ . (ii) If there are no other posts containing  $r_1$ , the discussion is lost. More precisely, it is attached to a resource that no longer exists in any post. While this is unpleasant, the other option – attaching the discussion to the new resource  $r_2$  – has two much more serious drawbacks: (i) If other users also had posts to resource  $r_1$ , after the update, the discussion would no longer be connected to the resources of their posts. (ii) This option would provide an easy means to hijack positive reviews. One must simply create a post for some high quality publication that attracts positive ratings and then use the update function to change the publication into another (possibly completely different) publication, such as a paper that oneself is the author of. To avoid such misuse, in BibSonomy, after an update, discussions stick with the old resource.

Throughout this chapter, we have argued for and against several features of social peer review, among others regarding the transparency and fairness of the process, and we have discussed the questions of who should be allowed to add reviews. To make the rating functions in BibSonomy as socially compatible as possible, the following choices were implemented: All registered users can add reviews and comment on all resources that are visible to them. This particularly includes resources in public posts and excludes resources that have been posted only in private posts. For writing a review, it is not required to have a post of the respective resource in one's own collection. Thus, when a user has posted a publication and this publication has been reviewed, it is not implied that the post's owner and the reviewer are the same person. Reviews consist of free text and an optional rating. Comments consist only of free text. Per resource, a user can add at most one review, but arbitrarily many comments. Thus, each user can rate each resource only once. Ratings are realized as star ratings on a scale between zero and five stars. Users can always edit or completely delete reviews or comments they wrote. For each review or comment, users have several visibility options. They can make the review visible to all users (public), to none but themselves (private), to their friends, or to a user group (e.g., their own working group). Additionally (and independent from the visibility setting) a user can choose to submit a review or comment anonymously. Concretely, anonymously here means that the review is linked to the user account internally, but this connection is not revealed to any other user. The advantage is that the reviewers are protected by anonymity, but at the same time, they can still edit their own reviews. Moreover, in case of a complaint, the operators of BibSonomy can get in touch with the user and mediate between the protester and the reviewer without disclosing the latter's identity. Independent from the visibility restriction of a review, all ratings are used to compute and aggregated rating (the





(a) A post containing a resource with two ratings.



(b) Rating distribution and average rating of a resource.

Figure 9.1: A resource with two ratings in BibSonomy.

arithmetic mean) for a resource. Users can add comments in reply to a review or to other comments. Thus, reviews can be supplemented, or refuted. Furthermore, authors or other users can discuss with the reviewer. Users that have been flagged as spammers in BibSonomy can submit reviews, however, their ratings are not included in the aggregated rating and their contributions are not publicly displayed.

When a resource has received ratings or comments, next to each post that contains the resource, the aggregated rating is shown, together with a link to the resource's discussion page. The latter contains the (visible) reviews and comments as well as a rating distribution (illustrated in Figure 9.1).

## 9.6 Conclusion

The Web 2.0 provides its users with a manifold of means to contribute, to shape, and to organize content online. In that context, rating portals assume the role of forums where users post their opinions for the purpose of judging other persons or products. Thus, such systems exist in a field of conflict between protective laws, namely the freedom of opinion and personality rights. Regarding our three research questions, we saw that online rating in general, as well as social peer review, including the selection of suitable peers, bears risks and opportunities, depending on the way these systems are designed.

In our section about online ratings in general, we have discussed the design of features regarding the creation of closed user groups, the mode and the aggregation of ratings, and their use in search engines. Generally, rating systems – even those for rating certain attributes of individuals – provide utility in a form that is compatible with (German) law (e.g., the spickmich decision). However, cases of alleged misuse or features that bear the potential to threaten personal rights, must be considered individually. It lies in the responsibility of the rating portals' operators to design their systems as socially compatible as possible and to react to complaints.

As a special case of online rating, we have considered social peer review, thus the case where researchers support their colleagues by providing their opinion on publications they have read. We have also compared social peer review to other forms of scholarly publication evaluation. In contrast to arbitrary rating systems, here, an additional obstacle is the identification of actual peers, users who are qualified to judge the merit of research. From the judicial point of view, it seems appropriate to consider all those who hold the freedom of sciences, as peers. However, to avoid hindering the social participation aspect of the Web 2.0, it is not required to actively restrict the circle of potential reviewers to experts. Quality control has to be realized through the technical design of the system, and we have proposed different means of assessing the expertise of potential reviewers or of weighting reviews according to their author's expertise. The choice of the most suitable means also depends on the goal of the system: For instance, it might or might not be desirable to include the opinions of practitioners outside the research community for assessing the value of a publication.

Another aspect of scholarly reviewing that is different from other rating systems, is that they have to provide strong incentives to motivate their users to actively contribute reviews. While in many contexts (e.g., movies or books), voicing one's own opinion is often enough reason for many users to contribute, for publicly writing critical reviews of colleagues' research, it takes courage as well as time and expertise. A possible incentive could be explicit prestige in the form of awards or reputation scores. Preferably, such a social peer-reviewing system would be run by an already respected organization within the research community. Further, integrating a reviewing system with those tools that researchers use for their literature work, like scholarly tagging systems, or with social networks, could lower the barrier for contribution.

### 9.6.1 Future Research

With the fast-paced development of web technology, it will become necessary to adapt the respective laws on the one hand, and the technology for protecting raters and those who are rated on the other hand. Moreover, research is conducted all over the world and in international collaboration. Thus, a successful social peer-reviewing system must address all researchers and would therefore have to be compliant with the laws in the respective countries (not only with German law, which we considered here). Finally, we have listed various ideas for realizing a social peer-reviewing system. The next step would be to test some of them in practice and to evaluate their success. A first step in that direction has been undertaken with the implementation of social peer review in BibSonomy.

## Chapter 10

### Conclusion and Outlook

In this thesis, we have contributed to the analysis of research, to the investigation of social bookmarking systems, and to the field of recommendations in folksonomies. In the face of information overload through the ever faster growing plethora of scholarly publications, researchers can be supported through analyses of their research communities and the respective publication corpora, as well as through systems that allow them to discover and to handle literature.

Focusing on data from the four phases of the life cycle of scholarly publications, we have created and demonstrated analyses that can support researchers. Using metadata from the creation phase of the publication life cycle, we have conducted analyses and visualizations of a research community, allowing its members to explore and inspect it from different angles. Through data from the dissemination phase, we could analyze the interactions of researchers during a conference. Such investigations can support conference organizers in evaluating the success of their event. Participants can better understand their own role in the community, and they can profit from tools like Conferator that guide users during conferences.

Data from the usage phase played the main role in this thesis. We have obtained it from a scholarly social bookmarking system (BibSonomy) and used it to analyze the system and to propose improvements. We have inspected the usage of the core features (tagging and retrieving) and improved recommender systems that support users by suggesting relevant literature. We have used data from the citation phase of the publication life cycle and compared it to usage data. We found that usage events in BibSonomy, like posts, exports, or views, bear a certain predictive power over future (citation-based) impact but also yield diverse measures. In social bookmarking systems, users can be supported by highlighting heavily used publications, indicating their possible relevance. Finally, we have discussed various design choices for the implementation of a social peer-reviewing system in a legal and socially compatible manner. In such a system, researchers can support each other by publishing their thoughts on scholarly articles.

In the remainder of this final chapter, we look ahead on future perspectives in those fields that we have contributed to in this thesis. While we have already pointed to specific opportunities for future work in the conclusion sections of each chapter, here, we take a rather broader perspective, speculating about possible directions for continuing research in these areas.

## 10.1 Analysis of Research Fields and Research Communities

In the past, investigations of research fields and research communities have predominantly been conducted by scientists in the form of published (thus, static) analyses. For example, the contributions of Chapters 3 and 4 fall into that category. Such studies are useful for those who are members of such communities or want to join them. Publishing such static analyses has two disadvantages: Due to the fast-paced development (exponential growth of the number of publications), such analyses are quickly outdated. Moreover, the specific field of research and the methodology to analyze it are chosen by those who conduct the analysis, rather than by their recipient.

One big next step would be providing flexible means for analyzing research fields and communities to all researchers on demand. Ideally, there would be an automated tool, a *scholarly explorer*, in which users can choose from the plethora of existing methods of analysis to explore research fields and research communities on their own and from different perspectives. Such a tool would give researchers the ability to track the status of their research community. It would help understand the influence of individual researchers or publications on specific fields and subfields of research. It would allow setting own work in relation to that of others.

By giving users the freedom of determining the data on which investigations are conducted and the freedom of selecting the method and of adjusting its focus, the scholarly explorer could produce its results based on the preferences of the analysis's recipient (instead of having an author decide for a reader). In short, it could be used to explore research following one's own interests at any point in time and based on one's current situation. The latter is important considering that different use cases require different analyses and different modes of presentation. For example, a user attending a research conference would probably want to see different visualizations or statistics (e.g., broad overviews and information that is quick to grasp) than students who are about to choose a direction for their thesis and thus can afford spending some time, trying different analyses, zooming in and out of different communities, and so on.

Until now, only few means of investigating research have been made available for everyone to use. For instance, Google Scholar supports the comparison of authors through three author-level metrics. Building on Google Scholar, the tool *Publish-or-Perish*<sup>1</sup> provides further statistics to compare selected authors. Another example is the Explorer from Altmetric<sup>2</sup> which presents aggregated data on scholarly publications from social networks and overviews on where the respective activities in the social media came from. Albeit useful, these tools have a restricted focus on particular data sources and the analyses they allow are rather rankings of entities regarding their impact using a fixed set of metrics, than actually analyzing a research field from various perspectives.

---

<sup>1</sup><http://www.harzing.com/resources/publish-or-perish>

<sup>2</sup><http://www.altmetric.com/products/explorer-for-institutions/>

What are the obstacles to building such a scholarly explorer as described above? It would be relatively straightforward to implement scientometrics methods, such as those in Chapters 3 or 4, in a tool. The bottleneck is rather the availability of suitable data. Acquiring the data from respective providers is often expensive and usually providers themselves have only a particular subset of publications (e.g., a catalog of editorially selected journal articles, like in the Web of Science, or publications and references that can be found on the web, like Google Scholar). Moreover, the data is often faulty, in part due to the method of collection (e.g., crawling from the web, extracting references from PDF) and in part due to errors in the publication metadata as provided by the publishers and authors.

Such errors might have only little influence on large overall studies, like comparing publication outputs of two countries. However, on a smaller scale, such as the level of individual authors or publications, they can have a strong influence.<sup>3</sup> Even big players, like Google Scholar struggle with the cleanness of the data and thus the validity of the presented statistics (e.g., per author) is questionable. However, cleaning the data is difficult and requires manual work.<sup>4</sup> Next to actual errors, other difficulties like ambiguity (e.g., two authors with the same name) or synonymy (e.g., authors changing their names in marriage) have to be tackled.

However, progress regarding an easier identification of authors and publications is being made. In Section 9.5, we have already mentioned DOI and ORCID. In the DOI system, publications receive a unique identifier that points to an online representation of the document. There, metadata of the document is stored and can be updated or corrected without changing the DOI. Similarly, ORCID provides a means to uniquely identify authors. Authors can register their own work with ORCID or have publishers submit information about new publications to ORCID for them. Using such unique identifiers bears the potential of easier automated data collection. Thus, it presents a promising approach to solve the above mentioned data bottleneck. Basically, a publication with authors and a reference section can be modeled as one DOI with a set of ORCIDs and references to a set of other DOIs. Error-prone tasks like parsing references and matching individual entities would be unnecessary, since interpreting references by following DOIs always leads to the same (official) metadata of a cited publication instead of to the version that the authors of the citing publication used. These developments foster the creation of automated data analysis tools, like the above mentioned scholarly explorer. Assuming that the data is made available (initiatives like the Microsoft Academic Graph<sup>5</sup> indicate progress in that direction), many analyses

---

<sup>3</sup>For example, misspelling an author's name can mean that this author is represented as two different entities in the data, which is both confusing and misleading as both entities will have lower indexes (e.g., publication counts) than the author would have when represented correctly.

<sup>4</sup>In the preparation of the dataset for the analyses in Chapter 3, despite the relatively small corpus, we had to correct astonishingly many publications and references, due to abbreviated or slightly altered titles, missing authors in references, various spellings of the same author, author names without first names or initials, and so on.

<sup>5</sup><http://research.microsoft.com/en-us/projects/mag/>

and visualizations can be automated and can be made available within a tool. Due to these promising prospects, we allow ourselves a longer glimpse into what will be possible provided the availability of the data.

- Researchers can be represented through variety of data: through their publications, their co-authors, or a set of metrics (which the user chooses), resulting from citations, altmetrics, reviews and ratings, or the analysis of scholarly networks.
- Numerical assessments would only be one part of the scholarly data explorer. It would cover other aspects, like relations between entities (e.g., collaborations between authors or influence between publications or between authors or groups of authors) and qualitative assessments, such as those from social peer review. We have already sketched a social peer-reviewing system in Section 9.5, which would provide a potential source for ratings and reviews that the scholarly explorer could build on.
- Users can be given the choice of the analysis and its parameters. In Chapter 3, we have shown a variety of visualizations and analyses of different aspects of the community of formal concept analysis. Depending on the direction into which users wish to explore a research field, they might choose a visualization (e.g., a map) of the community with its collaborations, of influences between subcommunities, of the developments of topics, or rankings of authors or publications based on some metric. By adjusting the parameters, users could, for example, select a zoom level in a visualization, thresholds to filter rare elements, and so on. For example, our method to visualize influences of particular papers or authors on subcommunities (Figures 3.6 and 3.7) required the setting of a threshold. Users could experiment with such parameters and compare the results for several choices.
- Users can select different restrictions of the datasets that are used in an analysis. Depending on the publication culture or on the user's preferences, in the analysis of publication metadata, the source of the data to conduct the analyses on should be variable. Data could be restricted by the publication venue, for example, only journal articles, only publications from some pre-defined catalog, only publications from particular venues, only publications from venues with a particular ranking, all peer-reviewed publications, or simply all publications together. Another restriction could be set by topic, for example, choosing publications by keyword, by search query, or by category of some literature classification system. In analyses relying on usage data from the social web, users could specify the systems from which data should be included or a time-frame in which the usage events have occurred.
- Next to allowing the users to influence the settings of an analysis, the scholarly explorer could also personalize other aspects of the results. Recommender systems

already assess the relevance of a document specifically for the user for whom the recommendations are computed rather than following some global relevance measure. Similarly, the selection of data can be personalized or recommendations can be provided based on the researchers or research communities a user showed interest in.

- All data can be dynamically updated. Digital publication catalogs already provide dynamically updated indexes. However, the published analyses of research fields are – naturally – static. Since DOIs are fixed identifiers even when metadata is changed (updated or corrected) and since in an automated tool, analyses can be computed just in time, the results can always reflect the most recent state of the available data.
- As a practical side-effect, the data of the scholarly explorer could be used to automatically generate and update bibliographies of different levels – for individual researchers, for departments, or for universities or other research facilities. Thus, researchers or institutions would be spared the effort of collecting and reporting. Moreover, if the data would officially serve as a representation of a researcher’s work, it would probably motivate many researchers to check the data about themselves and to correct errors or add missing entries, thus contribute to the validity of the collected data.

An exploration tool with these capabilities would greatly support the process of familiarizing one-self with a field of research. Next to the big challenge of overcoming the data bottleneck, other more methodological challenges must be addressed. Publications and researchers must be assigned to the research fields and subfields they belong to, such that users who want to explore such a community or field will find them included in their analysis. Moreover, offering automatic analysis to everyone comes with the caveat of leaving the interpretation of the presented figures, statistics, and relations to a possibly uniformed reader. In fact, many of today’s tools offer metrics (citation or publication counts, ranks, etc.) without further comment on their comparability or their significance. It will be a challenge to find ways to communicate these aspects in ways, users will recognize and accept, even without deeper knowledge of statistics or of the comparability. A requirement of technical nature would be to create the tool with open interfaces, such that new forms of analyses can be added easily.

Finally, the here described (hypothetical) scholarly explorer focuses on utilizing metadata. Even further utility could be provided by analyzing the actual publications, that is, their content. If that was available – the current activities to establish open access publishing aim at exactly that –, many more analyses could possibly be provided, from summaries of individual publications to automatically collected surveys on the current state of the art in a research field or to automatically compiled related work sections given a research paper draft as a query. By and large, the potential for providing utility to researchers through automated analyses of scholarly data is high and bears many opportunities for future research.

## 10.2 Social Bookmarking

The second main theme of this thesis is scholarly social bookmarking and we have investigated various aspects and extensions in Chapters 5 through 9. Particularly in our study of the usage of BibSonomy, we saw evidence that the principle of collecting resources for one's own purposes while profiting from the public collections of others seems to work. We can interpret that as a success of the idea of collaboratively collecting and annotating resources. At the same time, we can observe that over the last years, the market of social bookmarking systems has been consolidated, and many such systems have been discontinued,<sup>6</sup> among them the scholarly publication management system Connotea. On the other hand, existing tagging systems have extended their functionality beyond the tagging of resources through various features (recommendations, reviewing, discussions), and vice versa the practice of tagging has been included as a secondary feature – in many applications, probably most prominent in the form of hashtags in social media. In the following, we sketch two directions into which scholarly tagging systems, like BibSonomy, could be developed.

**Focusing on the Social Aspect.** One of the advantages of *social* bookmarking is that the collections of other users can be inspected and thus new resources can be discovered. Today's systems that feature social bookmarking have expanded the social dimension beyond this capability of discovering other users' content. In BibSonomy, CiteULike, and Mendeley, users can organize themselves in groups, interact in discussions, or share documents. Even further go systems like Academia.edu or ResearchGate, which promote themselves rather as social networks for researchers which also support bookmarking publications. They offer typical features, like maintaining a social network profile, following other users, endorsing others for skills, and so on. Specializing on academics, they allow presenting one's own publications, sharing and requesting documents, and even uploading drafts to gather feedback from colleagues. Particularly ResearchGate even computes its own impact scores for researchers. Even though the practice and especially its realization are questionable [Kraker and Lex, 2015], it can be a powerful motivator for researchers to become actively engaged in the network (to boost their scores and to compare with others). Scholarly social networks have successfully attracted huge numbers of users – Academia.edu reports more than 35 million members<sup>7</sup> and ResearchGate more than nine million.<sup>8</sup> Thus, extending the social component of a publication management system and adding more means for getting in touch and new incentives for contributing (scores, rankings, or achievements, by which users can compare to others) seems to be a promising development direction.

Systems that support tagging but also have a stronger social component than the plain collaborative collecting of publications, bear the opportunity of observing social interactions between researchers. Compared to our investigations of conference

---

<sup>6</sup>[https://en.wikipedia.org/wiki/List\\_of\\_social\\_bookmarking\\_websites](https://en.wikipedia.org/wiki/List_of_social_bookmarking_websites)

<sup>7</sup><https://www.academia.edu/> (accessed April 14, 2016)

<sup>8</sup><https://www.researchgate.net/about> (accessed April 14, 2016)



participants in Chapter 4, they allow the tracking of interactions over much longer time periods. Thus, they allow the study of collaborations and of research careers. Tagging systems with a stronger social component (or the other way around: social networks with a tagging feature) furthermore allow investigations on the connections between tagging and social interactions, for instance, the comparison of vocabularies in communities or of the diffusion of tags along the connections within the social network.

**Supporting the Creation Phase of the Scholarly Publication Life Cycle.** Scholarly bookmarking has made collecting and retrieving publication metadata easy and thus supports researchers when they assemble the literature sections of their own publications. However, the process of actually integrating these references into a draft, as well as the other steps of creating a publication are conducted elsewhere, outside the systems. Thus, a way to increase the utility of publication management systems would be to integrate the process of writing a publication. Collaborative writing software is available (e.g., the open-source software *sharelatex*<sup>9</sup>). Since scholarly bookmarking systems support the use and collection of literature, an integration of collaborative writing software would allow using the literature in references directly. Next to collaboratively writing also commenting and collaborative reviewing (e.g., of a draft before submission) could be integrated in publication management systems. BibSonomy already supports reviewing of publications and preprints, which could also be integrated further, for instance, to allow comments of specific text parts in draft.

Further possible extensions include the integration with submission and (pre-publication) reviewing systems, like EasyChair.<sup>10</sup> Thus, reviews could be written within the same system that the reviewer uses to manage literature and, which would facilitate the easy inclusion of references in the review. In turn it would be possible to store the review next to the reviewed submission. Thus, reviewers would have a collection of all the reviews they wrote. Finally, publication management systems could carry out the distribution of the final publication to other systems. For example with PUMA,<sup>11</sup> a BibSonomy-based publication management system for research institutions, it is already possible to forward publications to a local institutional publication repository.

For researchers, interesting opportunities for investigations would emerge from the data that is produced through the use of the above mentioned features. Publications could be studied during their creation, from draft to camera ready publication, possibly allowing the identification of different stages in the writing process or of writing styles of a publication's authors. Analyses of the edits or of deletions (in the text, or the inclusions or exclusions of publications in the references section) could give insights into topic shifts during the writing process or different preferences of different coauthors. It would also allow directly determining the usage events within the system (posts, views, exports, and so on) that lead to citations (as citations can be observed when a

---

<sup>9</sup><https://github.com/sharelatex/sharelatex>

<sup>10</sup><http://www.easychair.org/>

<sup>11</sup><http://www.academic-puma.de/>

publication is included as reference in a draft). As a result, recommender systems could provide support depending on the stage of the draft, including the recommendation of references for related work, keywords to describe the paper, or even of conferences or journals to submit the draft to.

### 10.3 Folksonomic Recommender Systems

In the field of folksonomic recommender systems, several continuations of the previous work suggest themselves. Following the lessons learned in Chapter 7, it would be reasonable to conduct a larger comparative study of folksonomic recommender algorithms. Several of the previously proposed algorithms should be tested in a robust benchmarking using several different cores. Impulses for new recommender strategies could come, among others, from the investigation of user behavior as documented in the user's posts and requests – in Chapter 5, we saw that we can distinguish users by their share of self-retrieval – or from the inclusion of further data – like rating data, for instance, from social peer review (Chapter 9) –, and we sketch such ideas in the following.

**Combining Rating and Folksonomy Data.** With the introduction of ratings and reviews into a tagging system, the folksonomy data is complemented with information on the personal impression of the quality of resources. By rating a resource, users reveal more explicitly whether they like it or not (in contrast to plain folksonomy data, where merely a user's act of posting a resource is interpreted as a sign that the user likes it). Rating data is the classical source for producing recommendations and thus such approaches can be applied in tagging systems where many users rate the resources. Furthermore, such approaches can be combined with those that rely on folksonomy data, thus exploiting both types of available data. For example, the proposed extensions of *FolkRank* in Chapter 8, could also be used to include rating data (e.g., by assigning weights based on ratings in the preference vector).

**Personalized Recommender Algorithm Selection.** We have shown that posts and requests paint different pictures on popularity (Chapters 5 and 6) of folksonomic entities. Moreover, we saw in Chapter 8 that different users might prefer different recommendation algorithms or different parametrizations of the same algorithm. Evaluating the performance of different recommender algorithms per user would allow to determine which users prefer which algorithms. Hereby, users could be described by features extracted from their behavior (based on posts and requests), among others their experience with using the system (e.g., using the number of previous requests as a simple heuristic), the share of self-retrieval, or a comparison between tags in posts and tags in retrieval. Other features could address tagging pragmatics, like those proposed by Körner et al. [2010], who showed that users can be distinguished into *categorizers* and *describers* by the way they use tags in their posts. Similar distinctions

are possible for pragmatics of using tags in retrieval. The challenge then would be to predict the algorithm that will most likely produce the best recommendations for the active user, based on that user's observed behavioral features. Considering that users can change their behavior during their time of using a system (e.g., gain experience), the features need to be updated regularly.

The result is a personalization of the choice of the recommendation algorithm, which is a special form of hybridization. In contrast to hybridization approaches like [Gemmell et al., 2012], there is no fixed combination of different algorithms, but rather a dynamic choosing of the most suitable candidate per user. For BibSonomy, we have added a preliminary study that shows the potential of choosing the algorithm per user, in Appendix E.

**Use-Case-based Recommendations.** Going one step further than selecting an algorithm per user would be to select the recommendation strategy based on user and use case. Let's consider the case of tag recommendations for scholarly publications. Users might prefer different tags for posting publications they authored than for posting publications they have read or will read. The tag recommendation strategy users would prefer, might also depend on their current work context. For instance, when users have stumbled upon a new resource while browsing the system, a different strategy might be appropriate than when users conduct literature research regarding a specific topic. In the latter case, a typical behavior pattern could be that users post publications quickly one after the other, storing publications they found for later reading. Thus it is reasonable to assume that recommending the same tags as for the previous posts (thus tags that are related to the current topic) should be successful. In contrast, in the case of stumbling onto an interesting publication and posting it, there is no indication that the current post is in any way related to a previous one.

This context of a post could be revealed by observing the previous activities of a user, which are recorded in the request logs of a bookmarking system. The research challenge would be to identify meaningful use cases, the respective indications in a user's requests and the most suitable recommendation algorithm that should be applied in each use case.

## 10.4 Further into the Future

Above, we have described perspectives for further research on these topics as well as for the development of new scholarly tools and extensions of the existing ones. Apart from these, each field will have to react and adapt to changes and new developments that arise, providing new challenges and new opportunities. The analysis of research has to react to the changing publication landscape resulting from various recent developments, such as open access journals and citations to preprints or other non-peer-reviewed material. In the liquid publications project [Casati et al., 2007] even more drastic changes have been proposed leading to a disentanglement from the traditional

publication system. Implementing their model would change the scholarly publication life cycle from the current model where publications are created as fixed texts that can be used and cited after their creation, to agile knowledge artifacts. Using ideas and processes from software engineering, liquid publications could be continuously developed, react to feedback from readers or include contributions from new authors.

Through social bookmarking systems, tagging has proven its worth, and at the latest through its use as a secondary feature in other systems, tagging has become an established technique in the social web. In each new system, tags might be used in new capacities, like we have seen happening in the social media (e.g., on Twitter, tags – as hashtags – are used for appending a new message to a discussion). Above, in Section 10.2, we have speculated about extended social tagging systems, where users can draft their publications collaboratively and discuss and review them. In such a system, tags could become the backbone for connecting the various entities, such as publications, text parts, comments, replies, edits, citations, and so on. One tag assignment would not only be a triple consisting of user, tag, and resource but instead could contain (next to user and tag) several resources of different types (e.g., a text part and a reference). Tags could also occur in notes that users take in documents, and these notes could be shared in the system and serve as descriptors for the resources they occur in. Such advances would call for new paradigms of using tags, as well as for new algorithms that recommend the tags or the resources that should be tagged.

Finally, the web is always changing, and new technologies come into focus, at the moment for example, mobile applications and ubiquitous computing. We can only be curious about the next big steps on the web, about how social tagging will fit in there, and how new developments can be used in scholarly applications to support researchers.

# Appendices



## Appendix A

### References of the Analyzed FCA Publications

This bibliography contains all those publications that are mentioned explicitly in the results of one of the analyses in Chapter 3. These publications belong either directly to the corpus that this study relies on, or to the references of some publication in the corpus.

#### References

- M. Barbut and B. Monjardet. *Ordre et classification: Algèbre et combinatoire*. Hachette, Paris, 1970. URL <http://www.worldcat.org/oclc/803811600>.
- G. Birkhoff. *Lattice Theory*. American Mathematical Society, Providence, 3rd edition, 1967. URL <http://www.worldcat.org/oclc/353879>.
- J. P. Bordat. Calcul pratique du treillis de galois d'une correspondance. *Informatiques et Sciences Humaines*, 96:31–47, 1986. URL <http://eudml.org/doc/94333>.
- C. Carpineto and G. Romano. *Concept Data Analysis: Theory and Applications*. John Wiley & Sons, Chichester, England, 2004. URL <http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470850558.html>.
- M. Chein and M.-L. Mugnier. Conceptual graphs: fundamental notions. *Revue d'Intelligence Artificielle*, 6(4):365–406, 1992. URL <http://www.lirmm.fr/~mugnier/ArticlesPostscript/RIA92ChMu.ps>.
- B. A. Davey and H. A. Priestley. *Introduction to lattices and order*. Cambridge University Press, Cambridge, 1990. ISBN 9780521784511. URL <http://www.cambridge.org/us/academic/subjects/mathematics/algebra/introduction-lattices-and-order-2nd-edition>.
- B. Ganter. Two basic algorithms in concept analysis. FB4-Preprint 831, TH Darmstadt, 1984. URL <http://www3.mathematik.tu-darmstadt.de/fb/mathe/preprints.html>.
- B. Ganter. Relational galois connections. In S. O. Kuznetsov and S. Schmidt, editors, *Formal Concept Analysis*, volume 4390 of *Lecture Notes in Computer Science*, pages 1–17. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-70828-5. doi:10.1007/978-3-540-70901-5\_1.

- B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, Berlin; New York, 1999. ISBN 3540627715 9783540627715. doi:10.1007/978-3-642-59830-2.
- B. Ganter, G. Stumme, and R. Wille, editors. *Formal Concept Analysis: Foundations and Applications*, volume 3626 of *Lecture Notes in Computer Science*. Springer, Berlin/Heidelberg, 2005. doi:10.1007/978-3-540-31881-1.
- J.-L. Guigues and V. Duquenne. Familles minimales d'implications informatives résultant d'un tableau de données binaires. *Mathématiques et Sciences Humaines*, 95:5–18, 1986. URL <http://eudml.org/doc/94331>.
- S. O. Kuznetsov and S. A. Obiedkov. Comparing performance of algorithms for generating concept lattices. *Journal of Experimental & Theoretical Artificial Intelligence*, 14(2-3):189–216, 2002. doi:10.1080/09528130210164170.
- D. Maier. *The Theory of Relational Databases*. Computer Science Press, Rockville, 1983. ISBN 0914894420. URL <http://www.worldcat.org/oclc/8306389>.
- M.-L. Mugnier and M. Chein. Représenter des connaissances et raisonner avec des graphes. *Revue d'Intelligence Artificielle*, 10(1):7–56, 1996. URL <http://www.lirmm.fr/~mugnier/ArticlesPostscript/RIA96MuCh.ps>.
- C. S. Peirce. *Collected Papers*. Harvard University Press, Cambridge, 1931. ISBN 9780674138001. URL <http://www.hup.harvard.edu/catalog.php?isbn=9780674138001>.
- S. Prediger. *Kontextuelle Urteilslogik mit Begriffsgraphen: Ein Beitrag zur Restrukturierung der mathematischen Logik*. Shaker, Aachen, 1998. ISBN 978-3-8265-3969-5. URL <http://www.shaker.de/de/content/catalogue/index.asp?ID=8&ISBN=978-3-8265-3969-5>.
- J. F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, Reading, 1984. URL <http://dl.acm.org/citation.cfm?id=4569>.
- J. F. Sowa. Conceptual graphs summary. In P. Eklund, T. Nagle, J. Nagle, and L. Gerholz, editors, *Conceptual structures: current research and practice*, pages 3–51. Ellis Horwood, 1992. ISBN 0-13-175878-0. URL <http://dl.acm.org/citation.cfm?id=168864>.
- J. F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks/Cole Publishing, Pacific Grove, 2000. ISBN 0-534-94965-7. URL <http://www.jfsowa.com/krbook/>.
- G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal. Computing iceberg concept lattices with TITANIC. *Data & Knowledge Engineering*, 42(2):189–222, 2002. ISSN 0169-023X. doi:10.1016/S0169-023X(02)00057-5.



- 
- R. Wille. Restructuring lattice theory: An approach based on hierarchies of concepts. In I. Rival, editor, *Ordered Sets*, pages 445–470, Dordrecht/Boston, 1982. Reidel. doi:10.1007/978-94-009-7798-3\_15.
- R. Wille. Conceptual graphs and formal concept analysis. In D. Lukose, H. Delugach, M. Keeler, L. Searle, and J. Sowa, editors, *Conceptual Structures: Fulfilling Peirce's Dream*, volume 1257 of *Lecture Notes in Computer Science*, pages 290–303. Springer, Berlin/Heidelberg, 1997. ISBN 978-3-540-63308-2. doi:10.1007/BFb0027878.



## Appendix B

### Correlations of Usage Metrics for Popular Topics in BibSonomy

For the sake of completeness, the following two tables contain the correlations between usage metrics in BibSonomy and citations in the same year ( $cit^{+0}$ ) and citations in the year after the usage ( $cit^{+1}$ ). Correlations are computed for the 30 most popular tags in BibSonomy (after stemming). The dataset and the details of the experiment are described in Chapter 6.

Table B.1: For each tag stem, the correlations (Pearson’s  $r$  and Spearman’s  $\rho$ ) between the behavioral metric and citations in the same year  $cit^{+0}$ . Correlations are computed over all publication-year pairs in which the publication has been used at least once (the according behavioral metric is non-zero).

tag stem	post		view		exp		expBib		req		tag	
	$r$	$\rho$	$r$	$\rho$	$r$	$\rho$	$r$	$\rho$	$r$	$\rho$	$r$	$\rho$
model	0.24	0.27	0.10	0.16	0.17	0.25	0.16	0.25	0.12	0.21	0.06	0.16
web	0.48	0.33	0.18	0.22	0.31	0.28	0.31	0.29	0.22	0.24	0.16	0.10
inform	0.36	0.39	0.14	0.25	0.52	0.31	0.54	0.30	0.23	0.28	0.13	0.12
network	0.42	0.32	0.15	0.24	0.23	0.27	0.23	0.27	0.18	0.24	0.15	0.14
system	0.40	0.26	0.11	0.17	0.16	0.23	0.15	0.23	0.12	0.18	0.09	0.15
semant	0.27	0.28	0.09	0.18	0.17	0.22	0.16	0.21	0.12	0.20	0.15	0.16
analysis	0.28	0.27	0.10	0.22	0.22	0.20	0.23	0.21	0.14	0.23	0.05	0.14
ontolog	0.34	0.31	0.13	0.19	0.26	0.29	0.27	0.28	0.17	0.21	0.13	0.13
social	0.28	0.31	0.06	0.18	0.10	0.26	0.10	0.26	0.07	0.21	0.06	0.11
commun	0.25	0.30	0.19	0.26	0.44	0.29	0.49	0.27	0.26	0.28	0.06	0.18
theori	0.26	0.20	0.23	0.24	0.56	0.26	0.56	0.26	0.35	0.26	0.04	0.18
comput	0.57	0.28	0.40	0.06	0.52	0.22	0.50	0.21	0.46	0.09	0.06	0.09
learn	0.29	0.33	0.08	0.15	0.14	0.26	0.15	0.29	0.12	0.19	0.00	0.02
manag	0.18	0.21	0.04	0.10	0.12	0.12	0.08	0.16	0.05	0.11	0.04	0.12
tag	0.38	0.34	0.18	0.23	0.35	0.38	0.37	0.39	0.22	0.25	0.33	0.28
design	0.45	0.29	0.38	0.14	0.64	0.18	0.65	0.21	0.47	0.15	0.05	0.13
evalu	0.20	0.20	0.15	0.12	0.34	0.17	0.34	0.18	0.20	0.16	0.05	0.08
folksonomi	0.34	0.35	0.10	0.27	0.21	0.39	0.21	0.39	0.13	0.30	0.23	0.36
softwar	0.31	0.16	0.15	0.18	0.42	0.19	0.44	0.20	0.21	0.17	0.15	0.02
collabor	0.46	0.37	0.16	0.23	0.28	0.32	0.27	0.30	0.20	0.27	0.15	0.14
data	0.25	0.22	0.17	0.29	0.23	0.25	0.22	0.25	0.25	0.30	0.12	0.07
knowledg	0.22	0.30	0.15	0.23	0.20	0.22	0.17	0.23	0.16	0.22	0.04	0.02
process	0.23	0.21	0.07	0.15	0.11	0.14	0.09	0.18	0.10	0.13	0.08	0.28
cluster	0.34	0.40	0.19	0.25	0.28	0.24	0.25	0.24	0.23	0.26	0.16	0.29
algorithm	0.32	0.26	0.25	0.26	0.17	0.33	0.17	0.34	0.20	0.27	0.03	0.02
web20	0.39	0.41	0.18	0.23	0.37	0.34	0.42	0.37	0.22	0.26	0.38	0.31
internet	0.25	0.21	0.11	0.11	0.13	0.23	0.15	0.27	0.13	0.13	0.16	0.04
search	0.32	0.27	0.13	0.12	0.25	0.20	0.26	0.22	0.06	0.14	0.23	0.15
structur	0.44	0.40	0.21	0.35	0.29	0.45	0.27	0.43	0.25	0.37	0.26	0.19
languag	0.20	0.27	0.10	0.15	0.21	0.23	0.19	0.23	0.13	0.15	0.02	0.09

Table B.2: For each tag stem, the correlations (Pearson’s  $r$  and Spearman’s  $\rho$ ) between the behavioral metric and citations in the following year  $cit^{+1}$ . Correlations are computed over all publication-year pairs in which the publication has been used at least once (the according behavioral metric is non-zero).

tag stem	post		view		exp		expBib		req		tag	
	$r$	$\rho$	$r$	$\rho$	$r$	$\rho$	$r$	$\rho$	$r$	$\rho$	$r$	$\rho$
model	0.25	0.26	0.10	0.18	0.17	0.26	0.16	0.26	0.12	0.22	0.06	0.15
web	0.49	0.38	0.20	0.29	0.31	0.29	0.31	0.30	0.23	0.29	0.16	0.08
inform	0.36	0.40	0.14	0.27	0.51	0.33	0.53	0.32	0.23	0.29	0.13	0.09
network	0.43	0.36	0.16	0.27	0.23	0.27	0.23	0.28	0.19	0.26	0.15	0.12
system	0.39	0.29	0.11	0.20	0.16	0.28	0.15	0.27	0.12	0.20	0.09	0.13
semant	0.30	0.31	0.14	0.25	0.20	0.21	0.19	0.21	0.16	0.24	0.15	0.16
analysi	0.32	0.28	0.11	0.25	0.21	0.22	0.22	0.23	0.14	0.26	0.05	0.11
ontolog	0.35	0.31	0.15	0.25	0.26	0.28	0.27	0.28	0.19	0.27	0.13	0.13
social	0.29	0.38	0.07	0.24	0.09	0.27	0.10	0.27	0.08	0.25	0.06	0.11
commun	0.26	0.30	0.19	0.29	0.43	0.29	0.48	0.27	0.26	0.32	0.06	0.17
theori	0.25	0.20	0.24	0.27	0.56	0.26	0.56	0.25	0.34	0.28	0.04	0.17
comput	0.58	0.28	0.39	0.11	0.55	0.22	0.52	0.21	0.45	0.13	0.06	0.05
learn	0.27	0.34	0.09	0.18	0.14	0.24	0.14	0.27	0.12	0.20	-0.00	-0.01
manag	0.17	0.22	0.04	0.12	0.11	0.17	0.08	0.20	0.05	0.15	0.03	0.07
tag	0.42	0.40	0.22	0.30	0.30	0.40	0.32	0.41	0.24	0.30	0.33	0.35
design	0.46	0.29	0.39	0.15	0.60	0.17	0.62	0.18	0.47	0.15	0.04	0.11
evalu	0.22	0.23	0.18	0.18	0.34	0.22	0.34	0.23	0.23	0.20	0.05	0.08
folksonomi	0.34	0.41	0.12	0.35	0.20	0.38	0.20	0.38	0.14	0.35	0.21	0.34
softwar	0.31	0.19	0.15	0.21	0.39	0.22	0.41	0.24	0.20	0.18	0.13	0.00
collabor	0.47	0.43	0.20	0.31	0.26	0.32	0.26	0.31	0.22	0.34	0.16	0.15
data	0.23	0.22	0.18	0.33	0.22	0.32	0.21	0.30	0.24	0.32	0.12	0.05
knowledg	0.20	0.25	0.15	0.24	0.19	0.21	0.16	0.22	0.15	0.23	0.04	0.00
process	0.24	0.22	0.10	0.18	0.18	0.13	0.14	0.18	0.13	0.15	0.08	0.27
cluster	0.34	0.41	0.19	0.28	0.28	0.25	0.25	0.25	0.23	0.29	0.17	0.30
algorithm	0.32	0.25	0.26	0.28	0.18	0.36	0.18	0.36	0.20	0.29	0.03	-0.02
web20	0.42	0.45	0.21	0.32	0.33	0.37	0.38	0.39	0.24	0.32	0.36	0.29
internet	0.26	0.22	0.14	0.14	0.14	0.27	0.16	0.31	0.15	0.14	0.12	-0.02
search	0.35	0.32	0.16	0.21	0.25	0.24	0.26	0.24	0.08	0.20	0.23	0.15
structur	0.44	0.40	0.20	0.42	0.28	0.47	0.26	0.45	0.24	0.43	0.26	0.18
languag	0.21	0.29	0.11	0.19	0.21	0.18	0.19	0.20	0.14	0.18	0.02	0.09



## Appendix C

### Tag Recommender Results on Different Cores

The following two figures are similar to Figure 7.4 in Chapter 7. Instead of the evaluation metric  $\text{pre}@5$ , here,  $\text{rec}@5$  and MAP are evaluated.

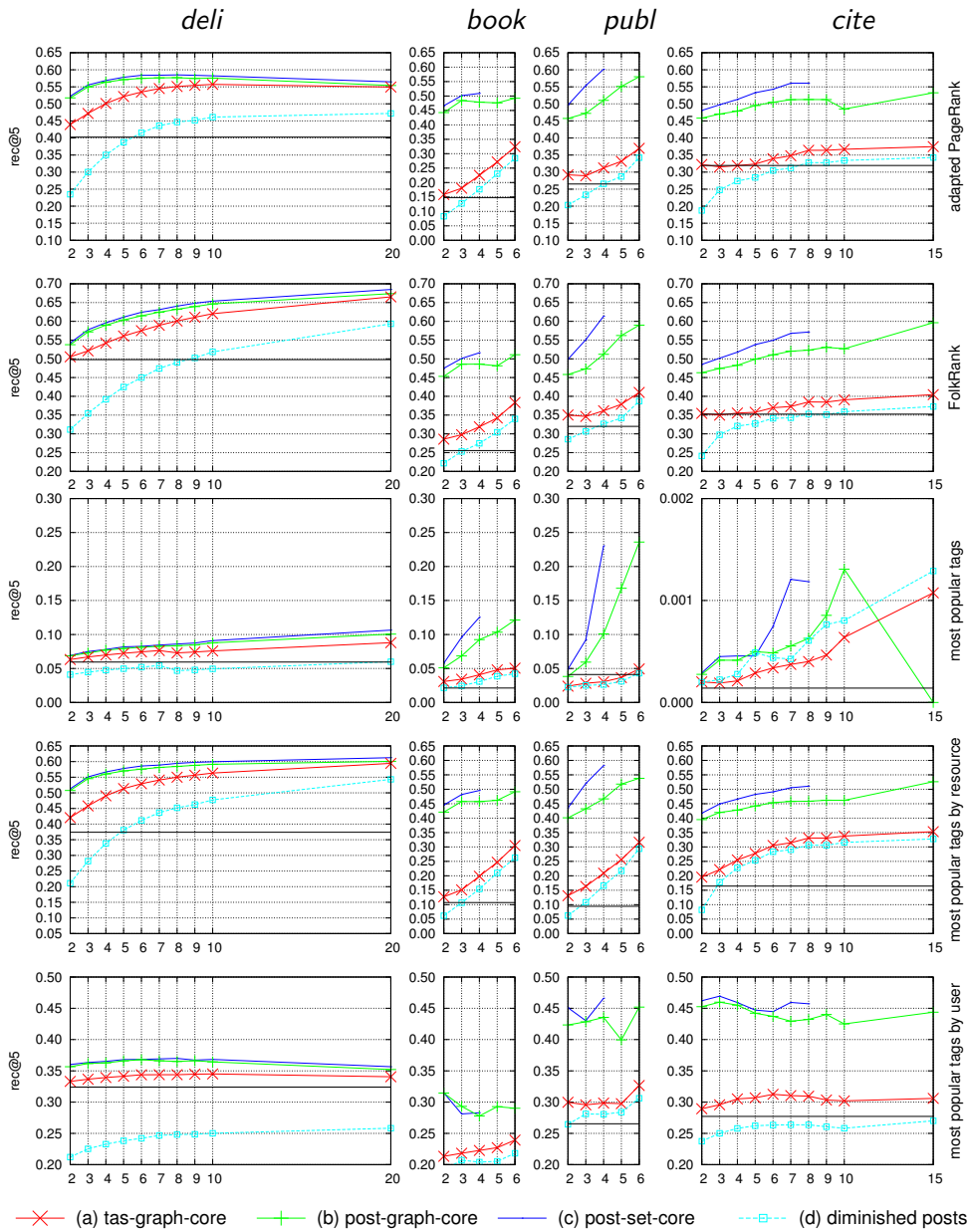


Figure C.1: The  $\text{rec}@5$  scores (on the  $y$ -axis) over the core level  $l$  (on the  $x$ -axis) for *deli*, *book*, *publ*, and *cite* for the five recommenders using modifications of *LeavePostOut*. Each column of plots represents the dataset specified at the top, each row contains results for the algorithm specified at the right, respectively. The horizontal lines depict the  $\text{rec}@5$  value for the respective raw dataset.



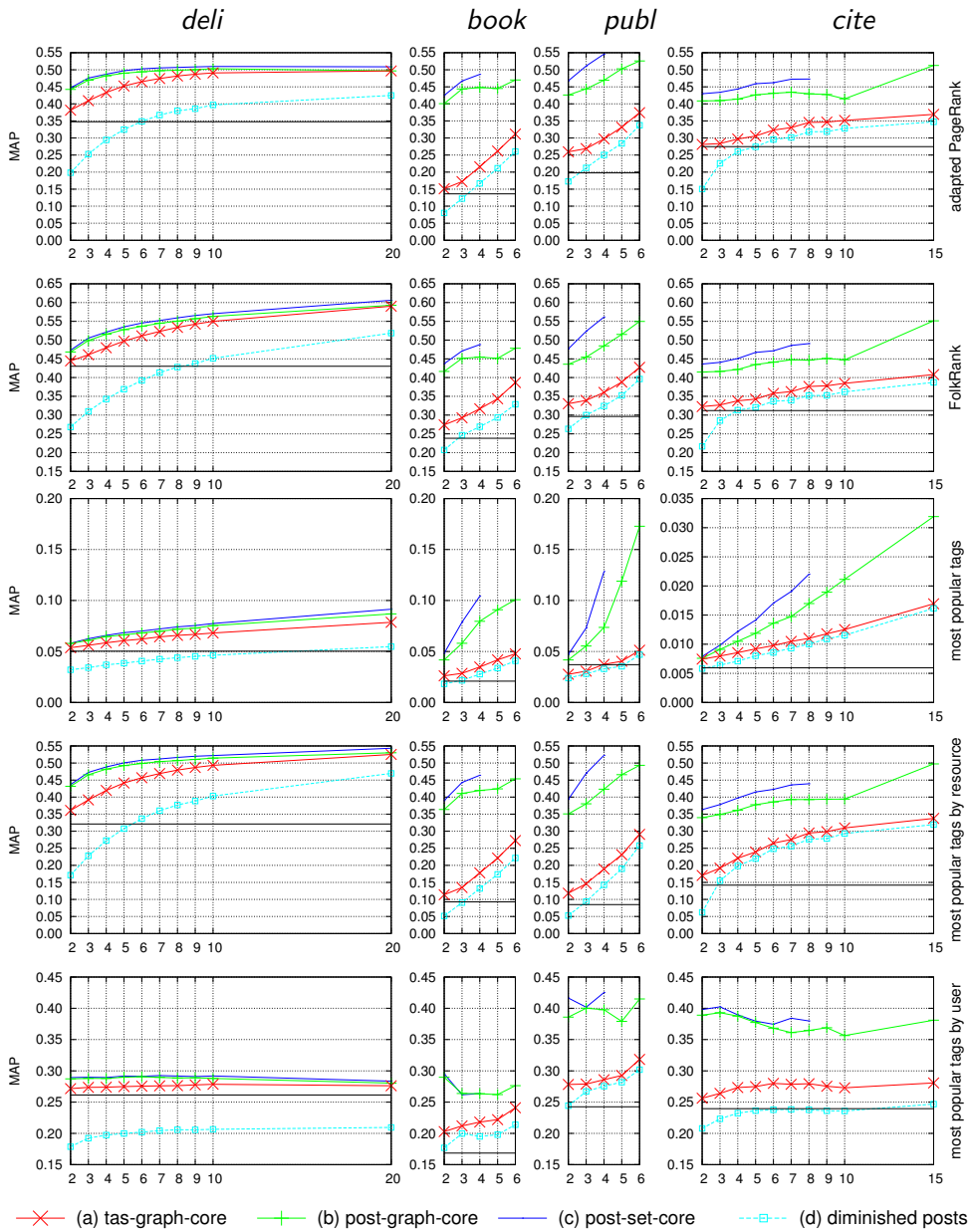


Figure C.2: The MAP scores (on the  $y$ -axis) over the core level  $l$  (on the  $x$ -axis) for *deli*, *book*, *publ*, and *cite* for the five recommenders using modifications of *LeavePostOut*. Each column of plots represents the dataset specified at the top, each row contains results for the algorithm specified at the right, respectively. The horizontal lines depict the MAP value for the respective raw dataset.



## Appendix D

### Results from Matrix Theory

The following three results are used in the proof sketch for the convergence of *FolkRank* in Section 8.2.3. They are taken from [Golub and Loan, 1996] and included here for the sake of self-containedness.

**Lemma D.1** (cf. Lemma 2.3.3 in [Golub and Loan, 1996]). *If  $F \in \mathbb{R}^{n \times n}$  and  $\|F\|_p < 1$ , then  $I - F$  is nonsingular and*

$$(I - F)^{-1} = \sum_{k=0}^{\infty} F^k$$

with

$$\|(I - F)^{-1}\|_p \leq \frac{1}{1 - \|F\|_p}.$$

**Theorem D.2** (Gershgorin Circle Theorem, cf. Theorem 7.2.1 in [Golub and Loan, 1996]). *If  $X^{-1}AX = D + F$  where  $D = \text{diag}(d_1, \dots, d_n)$  and  $F$  has zero diagonal entries, then*

$$\lambda(A) \subseteq \bigcup_{i=1}^n D_i,$$

where  $D_i = \left\{ z \in \mathbb{C} \mid |z - d_i| \leq \sum_{j=1}^n |f_{ij}| \right\}$ .

**Theorem D.3** (cf. Theorem 10.1.1 in [Golub and Loan, 1996]). *Suppose  $b \in \mathbb{R}^n$  and  $A = M - N \in \mathbb{R}^{n \times n}$  is nonsingular. If  $M$  is nonsingular and the spectral radius of  $M^{-1}N$  satisfies the inequality  $\rho(M^{-1}N) < 1$ , then the iterates  $x^{(k)}$  defined by  $Mx^{(k+1)} = Nx^{(k)} + b$  converge to  $x = A^{-1}b$  for any starting vector  $x^{(0)}$ .*



## Appendix E

### Personalized Recommender Algorithm Selection

One observation in Chapter 8 was that different users might prefer different recommender algorithms. As a consequence, a recommender system could choose a recommender algorithm (or its parameters) depending on the active user and, thus, improve the overall quality of the recommendations. For folksonomic recommender systems such approaches have previously been ignored; benchmarks of algorithms have usually been conducted by applying each algorithm to the same set of users. In the following, we discuss the idea of personalizing the choice of the recommender algorithm, and we present preliminary observations in BibSonomy to further substantiate this direction for future work.

In Chapter 5, we showed that users exhibit different behavior in their use of BibSonomy (e.g., how much they use the system or the share of requests that is spent on content of other users). Such usage features might be suitable to assign “the right recommender to the right user”. Kern et al. [2010] already noted that measures of tagging pragmatics (namely categorizers and describers) could serve as discriminating features for choosing the best recommendation algorithm per user. We add to their argument and propose usage data as another source for such features, based on the following preliminary study: We compute the performance of various recommender algorithms per user, as well as several usage features based on a user’s requests. Then, we determine correlations between the quality scores of each recommendation algorithm and each of the usage features. We show that different algorithms have different usage features with which they are correlated the most. This suggests that such usage features could be suitable for determining the recommendation algorithm that is chosen to compute recommendations for a user.

**Setup.** Using the BibSonomy dataset from Chapter 5, containing the logged requests of all users, we determine very simple usage features for each user. Basically, we count certain types of interactions, like the number of edits, the number of visits to other users, and so on. We compare these statistics to the tag recommender scores from Chapter 7 on the two BibSonomy datasets *publ* and *book*. We consider all users that are confirmed non-spammers, and we use the raw datasets (no core construction). Thus, for each user, we have their usage statistics and their recommendation results, and we can compute correlations.

Table E.1: Correlation between usage statistics (per user) and the precision scores of the five tag recommendation algorithms *FolkRank*, *adapted PageRank (APR)*, *most popular tags by user (by user)*, *most popular tags by resource (by resource)*, and *most popular tags (most popular)*. Reported is Spearman’s  $\rho$  for the datasets *publ* (upper) and *book* (lower). All correlation values are significant at the 0.05-level.

usage statistic	recommendations on <i>publ</i> , 3568 users				
	<i>FolkRank</i>	<i>APR</i>	<i>by user</i>	<i>by resource</i>	<i>most popular</i>
all requests	0.30	0.20	0.34	0.21	0.17
requests to users	0.32	0.22	0.36	0.23	0.18
requests to self	0.29	0.19	0.35	0.17	0.16
requests to others	0.32	0.26	0.29	0.37	0.20
requests to publications	0.32	0.23	0.36	0.25	0.19
requests to tags	0.27	0.18	0.29	0.21	0.16
distinct requested tags	0.24	0.15	0.26	0.21	0.15
edits	0.31	0.19	0.38	0.19	0.12

usage statistic	recommendations on <i>book</i> , 3205 users				
	<i>FolkRank</i>	<i>APR</i>	<i>by user</i>	<i>by resource</i>	<i>most popular</i>
all requests	0.26	0.18	0.34	0.22	0.14
requests to users	0.27	0.20	0.34	0.24	0.15
requests to self	0.25	0.18	0.35	0.21	0.15
requests to others	0.27	0.27	0.29	0.31	0.16
requests to tags	0.22	0.16	0.27	0.20	0.13
distinct requested tags	0.21	0.16	0.25	0.20	0.14
edits	0.26	0.15	0.38	0.19	0.12

**Analysis.** In Table E.1, we report Spearman’s  $\rho$  for correlations between the precision<sup>1</sup> scores of all five recommender algorithms from Chapter 7 and eight usage statistics, namely, the overall number of requests, as well as the numbers of requests to specific users, of requests to own content, of requests to content of others, of requests to publications,<sup>2</sup> of requests to tags, of different tags used in requests, and of edits (adding or updating posts).

The results for both datasets are roughly comparable. All correlations are small or moderate. We can, however, observe that the five different algorithms yield different

<sup>1</sup>Results for the recommendation quality metrics recall and mean average precision are qualitatively similar.

<sup>2</sup>We use this statistic only for *publ*. A similar statistic for bookmarks cannot be computed from the datasets. See Section 5.4 for details.

---

correlations with the various usage statistics: The lowest correlations are found between usage features and *most popular tags*. The results for *FolkRank* are relatively stable, between 0.29 and 0.32 on *publ* and between 0.25 and 0.27 on *book*. The exceptions are lower correlations for the two tag-related usage measures, which is rather surprising, since it is *tag* recommendations we evaluate. Correlations with *adapted PageRank* are strictly lower than those with *FolkRank* and closer to the results of *most popular tags*. The exception here is the number of requests to content of other users for which higher correlations are found.

The strongest correlations to the request counts yields *most popular tags by user*. Its highest correlation ( $\rho = 0.38$ ) is observed to the number of a user's edit requests (both datasets). Contrarily, for most of the other recommenders, the correlation with edits is rather lower than with other measures. Also, *most popular tags by user* is the only recommender that exhibits stronger correlations to the number of requests to oneself than with the number of requests to other users. This is consistent with intuition, since it seems to suggest that users who spend more time on their own collections also prefer their own previously used tags. Similarly, users who spend more requests on inspecting the collections of other users, might also prefer these users' tags, causing *most popular tags by resource* to be correlated much higher to the number of requests to others than to any of the other metrics.

**Conclusion.** By and large, for different recommender algorithms, we observe different correlations with various aspects of user behavior. This is evidence for the assumption that users who differ in the way they interact with the system also differ in their preference of recommendations. Thus, investigating selection strategies to assign different recommender algorithms to different users, is a reasonable approach for future work to further improve the recommender performance in tagging systems. To tackle this task, several challenges have to be overcome. First, discriminating user features through which users can be described, must be identified. Since the usage measures above are relatively simplistic, there is room for improvements using various compositions of them and further measurable aspects of usage behavior. A second challenge is the selection of suitable recommender algorithms. Many algorithms have been proposed (see Sections 7.3 and 8.3) and all of them could be checked for correlations with various usage properties. Eventually, there is the task of learning which algorithm to select for each (type of) user. Next to choosing one algorithm out of a set of candidates and assigning it to a user, one could also use a hybrid approach that employs a weighted combination of several algorithms. The personalization task in that scenario would be to choose these weights for each user.

Finally, the process of selecting the algorithm to produce the recommendations could also consider other factors, additionally to the active user. McNee et al. [2006] found that the algorithm (or rather the resulting list of recommendations) that a user prefers, depends on the task that a user wants to perform. In their user study, in the context of scholarly paper recommendation in a digital library, the tasks were, for

example, finding relevant papers that are closely related to the user's current research or finding interesting papers beyond the user's current research area. Systems like BibSonomy are used for different purposes, like collecting papers to read them later, managing cited literature, presenting own work (e.g., for reporting), or any mix thereof. Thus, it would be worth investigating how recommender systems can be chosen such that they fit to the user and to the task they try to accomplish.

Such studies require intensive investigation of user behavior, preferably on more than one tagging system. Moreover, through the large number of available candidate algorithms, large scale experiments will be necessary to find suitable selection strategies. Therefore, they are beyond the scope of this thesis and provide a promising direction for future research.



## Bibliography

- F. Abel, N. Henze, D. Krause, and M. Kriesell. On the effect of group structures on ranking strategies in folksonomies. In I. King and R. Baeza-Yates, editors, *Weaving Services and People on the World Wide Web*, pages 275–300. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-00569-5. doi:10.1007/978-3-642-00570-1\_14.
- F. Abel, N. Henze, R. Kawase, and D. Krause. The impact of multifaceted tagging on learning tag relations and search. In L. Aroyo, G. Antoniou, E. Hyvönen, A. ten Teije, H. Stuckenschmidt, L. Cabral, and T. Tudorache, editors, *The Semantic Web: Research and Applications*, volume 6089 of *Lecture Notes in Computer Science*, pages 90–105. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-13488-3. doi:10.1007/978-3-642-13489-0\_7.
- D. Abrams, R. Baecker, and M. Chignell. Information archiving with bookmarks: Personal web space construction and organization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '98*, pages 41–48, New York, NY, USA, 1998. ACM Press/Addison-Wesley Publishing Co. ISBN 0-201-30987-4. doi:10.1145/274644.274651.
- G. Adomavicius and J. Zhang. Impact of data characteristics on recommender systems performance. *ACM Transactions on Management Information Systems*, 3(1):3:1–3:17, Apr. 2012. ISSN 2158-656X. doi:10.1145/2151163.2151166.
- M. Agosti, F. Crivellari, and G. Di Nunzio. Web log analysis: A review of a decade of studies about information acquisition, inspection and interpretation of user interaction. *Data Mining and Knowledge Discovery*, 24(3):663–696, 2012. ISSN 1384-5810. doi:10.1007/s10618-011-0228-8.
- R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, SIGMOD '93*, pages 207–216, New York, NY, USA, 1993. ACM. ISBN 0-89791-592-5. doi:10.1145/170035.170072.
- R. Agrawal, M. Christoforaki, S. Gollapudi, A. Kannan, K. Kenthapadi, and A. Swaminathan. Mining videos from the web for electronic textbooks. In C. Glodeanu, M. Kaytoue, and C. Sacarea, editors, *Formal Concept Analysis*, volume 8478 of *Lecture Notes in Computer Science*, pages 219–234. Springer International Publishing, 2014. ISBN 978-3-319-07247-0. doi:10.1007/978-3-319-07248-7\_16.

- A. Ahmed, V. Batagelj, X. Fu, S.-H. Hong, D. Merrick, and A. Mrvar. Visualisation and analysis of the internet movie database. In *6th Int. Asia-Pacific Symposium on Visualization*, pages 17–24, 2007. doi:10.1109/APVIS.2007.329304.
- D. W. Aksnes. A macro study of self-citation. *Scientometrics*, 56(2):235–246, 2003. ISSN 0138-9130. doi:10.1023/A:1021919228368.
- H. Alani, M. Szomszor, C. Cattuto, W. Van den Broeck, G. Correndo, and A. Barrat. Live social semantics. In *The Semantic Web - ISWC 2009*, pages 698–714, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-04930-9. doi:10.1007/978-3-642-04930-9\_44.
- P. Albarrán and J. Ruiz-Castillo. References made and citations received by scientific articles. *Journal of the American Society for Information Science and Technology*, 62(1):40–49, 2011. ISSN 1532-2890. doi:10.1002/asi.21448.
- J. Alstott, E. Bullmore, and D. Plenz. powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE*, 9(1):1–11, 01 2014. doi:10.1371/journal.pone.0085777.
- J. I. Alvarez-Hamelin, L. Dall’Asta, A. Barrat, and A. Vespignani. K-core decomposition of internet graphs: Hierarchies, self-similarity and measurement biases. *Networks and Heterogeneous Media*, 3(2):371–393, 2008. ISSN 1556-1801. doi:10.3934/nhm.2008.3.371.
- M. Ames and M. Naaman. Why we tag: Motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI ’07*, pages 971–980, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-593-9. doi:10.1145/1240624.1240772.
- Y. An, J. Janssen, and E. E. Milios. Characterizing and mining the citation graph of the computer science literature. *Knowledge and Information Systems*, 6(6):664–678, Nov. 2004. ISSN 0219-3116. doi:10.1007/s10115-003-0128-3.
- R. Angelova, M. Lipczak, E. Milios, and P. Pralat. Characterizing a social bookmarking and tagging network. In *Proceedings of the Mining Social Data Workshop*, pages 21–25. ECAI, 2008. URL [http://www.math.ryerson.ca/~pralat/papers/2008\\_delicious.pdf](http://www.math.ryerson.ca/~pralat/papers/2008_delicious.pdf).
- M. Atzmueller, D. Benz, S. Doerfel, A. Hotho, R. Jäschke, B. E. Macek, F. Mitzlaff, C. Scholz, and G. Stumme. Enhancing social interactions at conferences. *it - Information Technology*, 53(3):101–107, 2011. doi:10.1524/itit.2011.0631.
- M. Atzmueller, M. Becker, S. Doerfel, M. Kibanov, A. Hotho, B.-E. Macek, F. Mitzlaff, J. Mueller, C. Scholz, and G. Stumme. Ubicon: Observing social and physical activities. In J. Bourgeois and A. Zomaya, editors, *Proceedings of the*

- 
- 2012 IEEE International Conference on Cyber, Physical and Social Computing, CPSCoM 2012, Besançon, France, 20-23 November, 2012, pages 317–324, Los Alamitos, CA, USA, 2012a. IEEE Computer Society. ISBN 978-1-4673-5146-1. doi:10.1109/GreenCom.2012.75.
- M. Atzmueller, S. Doerfel, A. Hotho, F. Mitzlaff, and G. Stumme. Face-to-face contacts at a conference: Dynamics of communities and roles. In M. Atzmueller, A. Chin, D. Helic, and A. Hotho, editors, *Modeling and Mining Ubiquitous Social Media - International Workshops MSM 2011, Boston, MA, USA, October 9, 2011, and MUSE 2011, Athens, Greece, September 5, 2011, Revised Selected Papers*, volume 7472 of *Lecture Notes in Computer Science*, pages 21–39. Springer Berlin Heidelberg, Heidelberg, Germany, 2012b. ISBN 978-3-642-33683-6. doi:10.1007/978-3-642-33684-3\_2.
- M. Atzmueller, M. Becker, M. Kibanov, C. Scholz, S. Doerfel, A. Hotho, B.-E. Macek, F. Mitzlaff, J. Mueller, and G. Stumme. UbiCon and its applications for ubiquitous social computing. *New Review of Hypermedia and Multimedia*, 20(1):53–77, 2014. doi:10.1080/13614568.2013.873488.
- M. Atzmueller, S. Doerfel, and F. Mitzlaff. Description-oriented community detection using exhaustive subgroup discovery. *Information Sciences*, 329:965–984, 2016a. ISSN 0020-0255. doi:10.1016/j.ins.2015.05.008.
- M. Atzmueller, A. Ernst, F. Krebs, C. Scholz, and G. Stumme. Formation and temporal evolution of social groups during coffee breaks. In *Big Data Analytics in the Social and Ubiquitous Context: 5th International Workshop on Modeling Social Media, MSM 2014, 5th International Workshop on Mining Ubiquitous and Social Environments, MUSE 2014, and First International Workshop on Machine Learning for Urban Sensor Data, SenseML 2014, Revised Selected Papers*, LNAI, pages 90–108. Springer Verlag, Heidelberg, Germany, 2016b. ISBN 978-3-319-29009-6. doi:10.1007/978-3-319-29009-6\_5.
- J. Bar-Ilan, S. Haustein, I. Peters, J. Priem, H. Shema, and J. Terliesner. Beyond citations: Scholars’ visibility on the social web. In É. Archambault, Y. Gingras, and V. Larivière, editors, *Proceedings of 17th International Conference on Science and Technology Indicators, Montréal: Science-Matrix and OST*, volume 1, pages 98–109, 2012. URL [http://2012.sticonference.org/Proceedings/vol1/Bar-Ilan\\_Beyond\\_98.pdf](http://2012.sticonference.org/Proceedings/vol1/Bar-Ilan_Beyond_98.pdf).
- A. Barrat, C. Cattuto, M. Szomszor, W. Van den Broeck, and H. Alani. Social dynamics in conferences: Analyses of data from the live social semantics application. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, editors, *The Semantic Web – ISWC 2010: 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part II*, volume 6497 of *Lecture Notes in Computer Science*, pages

- 17–33, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-17749-1. doi:10.1007/978-3-642-17749-1\_2.
- V. Batagelj and M. Zaveršnik. Generalized cores. arXiv:cs/0202039 [cs.DS], 2002. URL <http://arxiv.org/abs/cs/0202039>.
- V. Batagelj and M. Zaveršnik. Fast algorithms for determining (generalized) core groups in social networks. *Advances in Data Analysis and Classification*, 5(2): 129–145, 2011. ISSN 1862-5347. doi:10.1007/s11634-010-0079-y.
- M. Baur, M. Gaertler, R. Görke, M. Krug, and D. Wagner. Generating graphs with predefined k-core structure. In *Proceedings of the European Conference of Complex Systems*, 2007. URL <http://i11www.ira.uka.de/extra/publications/bggkw-ggpcs-07.pdf>.
- J. Beel, B. Gipp, S. Langer, and C. Breitingner. Research-paper recommender systems: A literature survey. *International Journal on Digital Libraries*, pages 1–34, 2015. ISSN 1432-5012. doi:10.1007/s00799-015-0156-0.
- F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference, IMC '09*, pages 49–62, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-771-4. doi:10.1145/1644893.1644900.
- D. Benz, A. Hotho, R. Jäschke, B. Krause, F. Mitzlaff, C. Schmitz, and G. Stumme. The social bookmark and publication management system BibSonomy. *The VLDB Journal*, 19(6):849–875, 2010a. ISSN 1066-8888. doi:10.1007/s00778-010-0208-4.
- D. Benz, A. Hotho, S. Stützer, and G. Stumme. Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge. In *Proceedings of the 2nd Web Science Conference (WebSci10)*, Raleigh, NC, USA, 2010b. URL <http://www.kde.cs.uni-kassel.de/pub/pdf/benz2010semantics.pdf>.
- M. Bianchini, M. Gori, and F. Scarselli. Inside pagerank. *ACM Transactions on Internet Technology*, 5(1):92–128, Feb. 2005. ISSN 1533-5399. doi:10.1145/1052934.1052938.
- L. Börneborn and P. Ingwersen. Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, 55(14):1216–1227, 2004. ISSN 1532-2890. doi:10.1002/asi.20077.
- V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. doi:10.1088/1742-5468/2008/10/P10008.
- T. Bogers. *Recommender Systems for Social Bookmarking*. PhD thesis, Tilburg University, Tilburg, The Netherlands, Dec. 2009. URL <http://ilk.uvt.nl/~toine/phd-thesis/>.

- 
- S. Bonzi and H. Snyder. Motivations for citation: A comparison of self citation and citation to others. *Scientometrics*, 21(2):245–254, 1991. ISSN 0138-9130. doi:10.1007/BF02017571.
- L. Bornmann. Do altmetrics point to the broader impact of research? an overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, 8(4):895–903, 2014a. doi:10.1016/j.joi.2014.09.005.
- L. Bornmann. Validity of altmetrics data for measuring societal impact: A study using data from altmetric and F1000Prime. *Journal of Informetrics*, 8(4):935–950, 2014b. doi:10.1016/j.joi.2014.09.007.
- L. Bornmann and H.-D. Daniel. What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80, 2008. doi:10.1108/00220410810844150.
- L. Bornmann and R. Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, pages 2215–2222, 2015. ISSN 2330-1643. doi:10.1002/asi.23329.
- Á. Borrego and J. Fry. Measuring researchers’ use of scholarly information through social bookmarking data: A case study of BibSonomy. *Journal of Information Science*, 38(3):297–308, 2012. doi:10.1177/0165551512438353.
- U. Brandes and T. Erlebach. Fundamentals. In U. Brandes and T. Erlebach, editors, *Network Analysis*, volume 3418 of *Lecture Notes in Computer Science*, pages 7–15. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-24979-5. doi:10.1007/978-3-540-31955-9\_2.
- L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 0885-6125. doi:10.1023/A:1010933404324.
- S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998. doi:10.1016/S0169-7552(98)00110-X.
- T. Brody, S. Harnad, and L. Carr. Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8):1060–1072, 2006. ISSN 1532-2890. doi:10.1002/asi.20373.
- B. C. Brookes. Biblio-, sciento-, infor-metrics??? what are we talking about? In L. Egghe and R. Rousseau, editors, *Informetrics 89/90: Selection of papers submitted for the second International Conference on Bibliometrics, Scientometrics, and Informetrics, London, Ontario, Canada, 5-7 July 1989*. Elsevier, 1990. URL <http://hdl.handle.net/1942/857>.

- C. H. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *Proceedings of the 15th international conference on World Wide Web, WWW '06*, pages 625–632, New York, NY, USA, 2006. ACM. ISBN 1-59593-323-9. doi:10.1145/1135777.1135869.
- M. Brzezinski. Power laws in citation distributions: Evidence from scopus. *Scientometrics*, 103(1):213–228, 2015. ISSN 0138-9130. doi:10.1007/s11192-014-1524-z.
- J. M. Campanario. Have referees rejected some of the most-cited articles of all times? *Journal of the American Society for Information Science*, 47(4):302–310, April 1996. ISSN 0002-8231. doi:10.1002/(SICI)1097-4571(199604)47:4<302::AID-ASI6>3.0.CO;2-0.
- I. Cantador, A. Bellogín, I. Fernández-Tobías, and S. López-Hernández. Semantic contextualisation of social tag-based profiles and item recommendations. In *E-Commerce and Web Technologies*, volume 85 of *Lecture Notes in Business Information Processing*, pages 101–113. Springer, Berlin/Heidelberg, 2011. ISBN 978-3-642-23014-1. doi:10.1007/978-3-642-23014-1\_9.
- M. J. Carman, M. Baillie, R. Gwadera, and F. Crestani. A statistical comparison of tag and query logs. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '09*, pages 123–130, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi:10.1145/1571941.1571965.
- F. Casati, F. Giunchiglia, and M. Marchese. Liquid publications: Scientific publications meet the web. Departmental Technical Report DIT-07-073, Information Engineering and Computer Science, December 2007. URL <http://eprints.biblio.unitn.it/1313/>.
- C. Cattuto, C. Schmitz, A. Baldassarri, V. D. P. Servedio, V. Loreto, A. Hotho, M. Grahl, and G. Stumme. Network properties of folksonomies. *AI Communications Journal, Special Issue on “Network Analysis in Natural Sciences and Engineering”*, 20(4):245–262, 2007. ISSN 0921-7126. URL <http://www.kde.cs.uni-kassel.de/stumme/papers/2007/cattuto2007network.pdf>.
- C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, and K. Thirunarayan, editors, *The Semantic Web - ISWC 2008 - 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings*, volume 5318 of *LNAI*, pages 615–631, Heidelberg, 2008. Springer. doi:10.1007/978-3-540-88564-1\_39.
- C. Cattuto, W. V. den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, and A. Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLoS ONE*, 5(7), 07 2010. doi:10.1371/journal.pone.0011596.

- 
- M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Transactions on Networking*, 17(5):1357–1370, 2009. ISSN 1063-6692. doi:10.1109/TNET.2008.2011358.
- H. Chun, H. Kwak, Y.-H. Eom, Y.-Y. Ahn, S. Moon, and H. Jeong. Comparison of online social relations in volume vs interaction: A case study of cyworld. In *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement*, IMC '08, pages 57–70, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-334-1. doi:10.1145/1452520.1452528.
- A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009. doi:10.1137/070710111.
- C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. ISSN 1573-0565. doi:10.1023/A:1022627411411.
- R. Costas, Z. Zahedi, and P. Wouters. Do “altmetrics” correlate with citations? extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, pages 2003–2019, 2014. ISSN 2330-1643. doi:10.1002/asi.23309.
- P. Cremonesi, Y. Koren, and R. Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 39–46, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-906-0. doi:10.1145/1864708.1864721.
- F. J. Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, Mar. 1964. ISSN 0001-0782. doi:10.1145/363958.363994.
- L. E. Damianos, D. Cuomo, J. Griffith, D. M. Hirst, and J. Smallwood. Exploring the adoption, utility, and social influences of social bookmarking in a corporate environment. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, HICSS '07, pages 86–95, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 0-7695-2755-8. doi:10.1109/HICSS.2007.219.
- C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *Proceedings of the 2nd ACM conference on Electronic commerce*, EC '00, pages 150–157, New York, NY, USA, 2000. ACM. ISBN 1-58113-272-7. doi:10.1145/352871.352889.
- J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006. ISSN 1532-4435. URL <http://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf>.

- R. Diestel. *Graph Theory*. Springer-Verlag Heidelberg, New York, 3 (electronic edition) edition, 2005. ISBN 978-3-642-14278-9. URL <http://diestel-graph-theory.com/>.
- S. Doerfel and R. Jäschke. An analysis of tag-recommender evaluation procedures. In *Proceedings of the 7th Conference on Recommender systems, RecSys '13*, pages 343–346, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2409-0. doi:10.1145/2507157.2507222.
- S. Doerfel, R. Jäschke, A. Hotho, and G. Stumme. Leveraging publication metadata and social data into folkRank for scientific publication recommendation. In *Proceedings of the 4th ACM RecSys workshop on Recommender systems and the social web*, pages 9–16, New York, NY, USA, 2012a. ACM. doi:10.1145/2365934.2365937.
- S. Doerfel, R. Jäschke, and G. Stumme. Publication analysis of the formal concept analysis community. In F. Domenach, D. Ignatov, and J. Poelmans, editors, *Formal Concept Analysis – 10th International Conference, ICFCA 2012, Leuven, Belgium, May 7-10, 2012. Proceedings*, volume 7278 of *Lecture Notes in Artificial Intelligence*, pages 77–95, Berlin/Heidelberg, May 2012b. Springer. ISBN 978-3-642-29892-9. doi:10.1007/978-3-642-29892-9\_12.
- S. Doerfel, A. Hotho, A. Kartal-Aydemir, A. Roßnagel, and G. Stumme. Empfehlungssysteme für wissenschaftliche Publikationen. In *Informationelle Selbstbestimmung im Web 2.0*, pages 113–148. Springer Berlin Heidelberg, 2013a. ISBN 978-3-642-38055-6. doi:10.1007/978-3-642-38056-3\_6.
- S. Doerfel, A. Hotho, A. Kartal-Aydemir, A. Roßnagel, and G. Stumme. *Informationelle Selbstbestimmung im Web 2.0 – Chancen und Risiken sozialer Verschlagwortungssysteme*. Xpert.press. Springer Berlin Heidelberg, 2013b. ISBN 978-3-642-38055-6. doi:10.1007/978-3-642-38056-3.
- S. Doerfel, A. Hotho, A. Kartal-Aydemir, A. Roßnagel, and G. Stumme. Online-Literaturbewertungssystem als Social-Peer-Review. In *Informationelle Selbstbestimmung im Web 2.0*, Xpert.press, pages 61–112. Springer Berlin Heidelberg, 2013c. ISBN 978-3-642-38055-6. doi:10.1007/978-3-642-38056-3\_6.
- S. Doerfel, D. Zoller, P. Singer, T. Niebler, A. Hotho, and M. Strohmaier. Of course we share! testing assumptions about social tagging systems. arXiv:1401.0629 [cs.IR], 2014a. URL <http://arxiv.org/abs/1401.0629>.
- S. Doerfel, D. Zoller, P. Singer, T. Niebler, A. Hotho, and M. Strohmaier. Evaluating assumptions about social tagging: A study of user behavior in BibSonomy. In T. Seidl, M. Hassani, and C. Beecks, editors, *Proceedings of the 16th LWA Workshops: KDML, IR and FGWM, Aachen, Germany, September 8-10, 2014*. CEUR-WS.org, 2014b. URL <http://ceur-ws.org/Vol-1226/paper06.pdf>.



- 
- S. Doerfel, D. Zoller, P. Singer, T. Niebler, A. Hotho, and M. Strohmaier. How social is social tagging? In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 251–252, Republic and Canton of Geneva, Switzerland, 2014c. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-2745-9. doi:10.1145/2567948.2577301.
- S. Doerfel, R. Jäschke, and G. Stumme. The role of cores in recommender benchmarking for social bookmarking systems. *ACM Transactions on Intelligent Systems and Technology*, 7(3):40:1–40:33, February 2016a. doi:10.1145/2700485.
- S. Doerfel, D. Zoller, P. Singer, T. Niebler, A. Hotho, and M. Strohmaier. What users actually do in a social tagging system: A study of user behavior in BibSonomy. *ACM Transactions on the Web*, 10(2):14:1–14:32, May 2016b. doi:10.1145/2896821.
- S. Duarte Torres, I. Weber, and D. Hiemstra. Analysis of search and browsing behavior of young users on the web. *ACM Transactions on the Web*, 8(2):7:1–7:54, Mar. 2014. ISSN 1559-1131. doi:10.1145/2555595.
- N. Eagle and A. S. Pentland. Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, 10:255–268, March 2006. ISSN 1617-4909. doi:10.1007/s00779-005-0046-3.
- F. Eisterlehner, A. Hotho, and R. Jäschke, editors. *ECML PKDD Discovery Challenge 2009 (DC09)*, volume 497 of *CEUR-WS.org*, Sept. 2009. URL <http://ceur-ws.org/Vol1-497>.
- D. Endres, R. Adam, M. Giese, and U. Noppeneay. Understanding the semantic structure of human fmri brain recordings with formal concept analysis. In F. Domenach, D. Ignatov, and J. Poelmans, editors, *Formal Concept Analysis*, volume 7278 of *Lecture Notes in Computer Science*, pages 96–111. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-29891-2. doi:10.1007/978-3-642-29892-9\_13.
- G. Eysenbach. Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research (JMIR)*, 13(4), 2011. doi:10.2196/jmir.2012.
- M. Falagas and V. Alexiou. The top-ten in journal impact factor manipulation. *Archivum Immunologiae et Therapiae Experimentalis*, 56(4):223–226, 2008. ISSN 0004-069X. doi:10.1007/s00005-008-0024-5.
- C. Ferguson, A. Marcus, and I. Oransky. Publishing: The peer-review scam. *Nature News*, 515(7528):480, Nov. 2014. doi:10.1038/515480a.
- M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine*

- Learning Research*, 15(1):3133–3181, Jan. 2014. ISSN 1532-4435. URL <http://jmlr.org/papers/v15/delgado14a.html>.
- S. Ferré. The efficient computation of complete and concise substring scales with suffix trees. In S. Kuznetsov and S. Schmidt, editors, *Formal Concept Analysis*, volume 4390 of *Lecture Notes in Computer Science*, pages 98–113. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-70828-5. doi:10.1007/978-3-540-70901-5\_7.
- F. Font, J. Serra, and X. Serra. Analysis of the impact of a tag recommendation system in a real-world folksonomy. *ACM Transactions on Intelligent Systems and Technology*, 7(1), 2016. doi:10.1145/2743026.
- S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5):75–174, 2010. ISSN 0370-1573. doi:10.1016/j.physrep.2009.11.002.
- L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977. ISSN 00380431. doi:10.2307/3033543.
- T. Fu, Q. Song, and D. Chiu. The academic social network. *Scientometrics*, 101(1):203–239, 2014. doi:10.1007/s11192-014-1356-x.
- E. R. Gansner and S. C. North. An open graph visualization system and its applications to software engineering. *Software Practice & Experience*, 30(11):1203–1233, Sept. 2000. doi:10.1002/1097-024X(200009)30:11<1203::AID-SPE338>3.3.CO;2-E.
- E. R. Gansner, Y. Hu, and S. G. Kobourov. GMap: Drawing graphs as maps. arXiv:0907.2585 [cs.CG], July 2009. URL <http://arxiv.org/abs/0907.2585>.
- B. Ganter. Two basic algorithms in concept analysis. FB4-Preprint 831, TH Darmstadt, 1984. URL <http://www3.mathematik.tu-darmstadt.de/fb/mathe/preprints.html>.
- B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, Berlin; New York, 1999. ISBN 3540627715 9783540627715. doi:10.1007/978-3-642-59830-2.
- E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479, Nov. 1972. doi:10.1126/science.178.4060.471.
- J. Gemmell, T. R. Schimoler, L. Christiansen, and B. Mobasher. Improving folkrank with item-based collaborative filtering. In D. Jannach, W. Geyer, J. Freyne, S. S. Anand, C. Dugan, B. Mobasher, and A. Kobsa, editors, *ACM RecSys'09 Workshop on Recommender Systems and the Social Web*, volume 532 of *CEUR-WS.org*, pages 17–24, Oct. 2009. URL <http://ceur-ws.org/Vol-532/paper3.pdf>.

- 
- J. Gemmell, T. Schimoler, B. Mobasher, and R. Burke. Resource recommendation in social annotation systems: A linear-weighted hybrid approach. *Journal of Computer and System Sciences*, 78(4):1160–1174, 2012. ISSN 0022-0000. doi:10.1016/j.jcss.2011.10.006.
- C. Giatsidis, D. M. Thilikos, and M. Vazirgiannis. Evaluating cooperation in communities with the k-core structure. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 87–93, 2011. doi:10.1109/ASONAM.2011.65.
- R. J. Glushko, P. P. Maglio, T. Matlock, and L. W. Barsalou. Categorization in the wild. *Trends in Cognitive Sciences*, 12(4):129–135, 2008. ISSN 1364-6613. doi:10.1016/j.tics.2008.01.007.
- W. Glänzel and H. Moed. Journal impact measures in bibliometric research. *Scientometrics*, 53(2):171–193, 2002. ISSN 0138-9130. doi:10.1023/A:1014848323806.
- D. Godoy and A. Corbellini. Folksonomy-based recommender systems: A state-of-the-art review. *International Journal of Intelligent Systems*, pages 314–346, 2015. ISSN 1098-111X. doi:10.1002/int.21753.
- S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006. doi:10.1177/0165551506062337.
- G. H. Golub and C. F. V. Loan. *Matrix Computations*. The Johns Hopkins University Press, third edition, 1996. ISBN 0-8018-5414-8. URL <https://jhupbooks.press.jhu.edu/content/matrix-computations>.
- J. González Calabozo, C. Peláez-Moreno, and F. Valverde-Albacete. Gene expression array exploration using  $\mathcal{K}$ -formal concept analysis. In P. Valtchev and R. Jäschke, editors, *Formal Concept Analysis*, volume 6628 of *Lecture Notes in Computer Science*, pages 119–134. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-20513-2. doi:10.1007/978-3-642-20514-9\_11.
- S. Greaves, J. Scott, M. Clarke, L. Miller, T. Hannay, A. Thomas, and P. Campbell. Nature’s trial of open peer review. *Nature International Weekly Journal of Science*, Web focuses, Science and politics, Peer Review: Debate, December 2006. doi:10.1038/nature05535.
- Z. Guan, C. Wang, J. Bu, C. Chen, K. Yang, D. Cai, and X. He. Document recommendation in social tagging services. In *Proceedings of the 19th International Conference on World Wide Web, WWW ’10*, pages 391–400, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi:10.1145/1772690.1772731.
- J.-L. Guigues and V. Duquenne. Familles minimales d’implications informatives résultant d’un tableau de données binaires. *Mathématiques et Sciences Humaines*, 95:5–18, 1986. URL <http://eudml.org/doc/94331>.

- M. R. Haley. Ranking top economics and finance journals using microsoft academic search versus google scholar: How does the new publish or perish option compare? *Journal of the Association for Information Science and Technology*, 65(5):1079–1084, 2014. ISSN 2330-1643. doi:10.1002/asi.23080.
- H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th International Conference on World Wide Web, WWW '07*, pages 211–220, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7. doi:10.1145/1242572.1242602.
- S. Haustein and I. Peters. Using social bookmarks and tags as alternative indicators of journal content description. *First Monday*, 2012. doi:10.5210/fm.v17i11.4110.
- S. Haustein and T. Siebenlist. Applying social bookmarking data to evaluate journal usage. *Journal of Informetrics*, 5(3):446–457, 2011. ISSN 1751-1577. doi:10.1016/j.joi.2011.04.002.
- S. Haustein, V. Larivière, M. Thelwall, D. Amyot, and I. Peters. Tweets vs. Mendeley readers: How do these two social media metrics differ? *it - Information Technology*, 56(5):207–215, September 2014a. ISSN 21967032. doi:10.1515/itit-2014-1048.
- S. Haustein, I. Peters, C. R. Sugimoto, M. Thelwall, and V. Larivière. Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *Journal of the Association for Information Science and Technology*, 65(4):656–669, 2014b. ISSN 2330-1643. doi:10.1002/asi.23101.
- T. Heck, I. Peters, and W. G. Stock. Testing collaborative filtering against co-citation analysis and bibliographic coupling for academic author recommendation. In J. Freyne, S. S. Anand, I. Guy, and A. Hotho, editors, *Workshop on Recommender Systems and the Social Web (ACM RecSys'11)*, 2011. URL <http://www.dcs.warwick.ac.uk/~ssanand/RSWeb11/rsweb2011proceedingsfinal.pdf>.
- M. Heckner, M. Heilemann, and C. Wolff. Personal information management vs. resource sharing: Towards a model of information behaviour in social tagging systems. In *Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, San Jose, CA, USA, May 2009. URL <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/212>.
- J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1): 5–53, 2004. ISSN 1046-8188. doi:10.1145/963770.963772.
- D. Hicks, P. Wouters, L. Waltman, S. de Rijcke, and I. Rafols. Bibliometrics: The leiden manifesto for research metrics. *Nature News*, 520(7548):429, Apr. 2015. doi:10.1038/520429a.

- 
- J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572, Nov. 2005. doi:10.1073/pnas.0507655102.
- S. Hornbostel and D. Simon. Einleitung: Wie viel (In-)Transparenz ist notwendig? Peer Review Revisited. In S. Hornbostel and D. Simon, editors, *Wie viel (In-)Transparenz ist notwendig? Peer Review Revisited*, volume 1, pages 5–6. Institut für Forschungsinformation und Qualitätssicherung, 2006. URL <http://nbn-resolving.de/urn:nbn:de:0168-ssoar-25300>.
- A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Trend detection in folksonomies. In *Proceedings of the First International Conference on Semantic and Digital Media Technologies*, SAMT'06, pages 56–70, Berlin, Heidelberg, 2006a. Springer-Verlag. ISBN 3-540-49335-2, 978-3-540-49335-8. doi:10.1007/11930334\_5.
- A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. BibSonomy: A social bookmark and publication sharing system. In A. de Moor, S. Polovina, and H. Delugach, editors, *Proceedings of the First Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, pages 87–102, Aalborg, Denmark, 2006b. Aalborg Universitetsforlag. ISBN 87-7307-769-0. URL <http://www.kde.cs.uni-kassel.de/stumme/papers/2006/hotho2006bibsonomy.pdf>.
- A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In Y. Sure and J. Domingue, editors, *The Semantic Web: Research and Applications: 3rd European Semantic Web Conference, ESWC 2006 Budva, Montenegro, June 11-14, 2006 Proceedings*, volume 4011 of *LNCS*, pages 411–426, Berlin/Heidelberg, 2006c. Springer Berlin Heidelberg. ISBN 978-3-540-34545-9. doi:10.1007/11762256\_31.
- A. Hotho, D. Benz, R. Jäschke, and B. Krause, editors. *ECML PKDD Discovery Challenge 2008 (RSDC'08)*, 2008. URL [http://www.kde.cs.uni-kassel.de/ws/rsdc08/pdf/all\\_rsdv2.pdf](http://www.kde.cs.uni-kassel.de/ws/rsdc08/pdf/all_rsdv2.pdf).
- M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM. ISBN 1-58113-888-1. doi:10.1145/1014052.1014073.
- P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket switched networks and human mobility in conference environments. In *Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, WDTN '05, pages 244–251, New York, NY, USA, 2005. ACM. ISBN 1-59593-026-4. doi:10.1145/1080139.1080142.
- L. Isella, M. Romano, A. Barrat, C. Cattuto, V. Colizza, W. Van den Broeck, F. Gesualdo, E. Pandolfi, L. Ravà, C. Rizzo, and A. E. Tozzi. Close encounters in a

- pediatric ward: Measuring face-to-face proximity and mixing patterns with wearable sensors. *PLoS ONE*, 6(2):1–10, 02 2011a. doi:10.1371/journal.pone.0017144.
- L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. V. den Broeck. What's in a crowd? analysis of face-to-face behavioral networks. *Journal of Theoretical Biology*, 271(1):166–180, 2011b. ISSN 0022-5193. doi:10.1016/j.jtbi.2010.11.033.
- S. Janson, T. Luczak, and A. Rucinski. *Theory of random graphs*. John Wiley and Sons, New York; Chichester, 2000. ISBN 0471175412 9780471175414. doi:10.1002/9781118032718.
- J. Jiang, C. Wilson, X. Wang, W. Sha, P. Huang, Y. Dai, and B. Y. Zhao. Understanding latent interactions in online social networks. *ACM Transactions on the Web*, 7(4): 18:1–18:39, Nov. 2013. ISSN 1559-1131. doi:10.1145/2517040.
- A. Jobmann, C. P. Hoffmann, S. Künne, I. Peters, J. Schmitz, and G. Wollnik-Korn. Altimetrics for large, multidisciplinary research groups: Comparison of current tools. *Bibliometrie – Praxis und Forschung*, 3, 2014. URL <http://www.bibliometrie-pf.de/article/view/205>.
- R. Jäschke, L. Balby Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In J. N. Kok, J. Koronacki, R. Lopez de Mantaras, S. Matwin, D. Mladenič, and A. Skowron, editors, *Knowledge Discovery in Databases: PKDD 2007: 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17-21, 2007. Proceedings*, volume 4702 of *LNCS*, pages 506–514, Berlin/Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-74976-9. doi:10.1007/978-3-540-74976-9\_52.
- R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in social bookmarking systems. *AI Communications*, 21(4):231–247, 2008. ISSN 0921-7126. doi:10.3233/AIC-2008-0438.
- R. Jäschke, F. Eisterlehner, A. Hotho, and G. Stumme. Testing and evaluating tag recommenders in a live system. In *RecSys '09: Proceedings of the third ACM Conference on Recommender Systems*, pages 369–372, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-435-5. doi:10.1145/1639714.1639790.
- R. Jäschke, A. Hotho, F. Mitzlaff, and G. Stumme. Challenges in tag recommendations for collaborative tagging systems. In *Recommender Systems for the Social Web*, volume 32 of *Intelligent Systems Reference Library*, pages 65–87. Springer, Berlin/Heidelberg, 2012. ISBN 978-3-642-25694-3. doi:10.1007/978-3-642-25694-3\_3.
- A. Jøsang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007. ISSN 0167-9236. doi:10.1016/j.dss.2005.05.019. Emerging Issues in Collaborative Commerce.

- 
- J. Kamp and K.-N. Peifer. Datenschutz und Persönlichkeitsrecht Anwendung der Grundsätze über Produktkritik auf das Bewertungsportal »spickmich.de«? *ZUM*, 53:185–190, 2009. URL <https://beck-online.beck.de/?typ=reference&y=300&b=2009&s=185&z=ZUM>.
- A. Kartal, S. Doerfel, A. Roßnagel, and G. Stumme. Privatsphären- und Datenschutz in Community-Plattformen: Gestaltung von Online-Bewertungsportalen. In H.-U. Heiß, P. Pepper, H. Schlingloff, and J. Schneider, editors, *Informatik 2011 - Informatik schafft Communities - Proceedings der 41. GI-Jahrestagung*, volume 192 of *Lecture Notes in Informatics*, page 412. Gesellschaft für Informatik e.V. (GI), Bonner Köllen Verlag, 10 2011. URL <http://www.informatik2011.de/541.html>.
- G. Karypis. Evaluation of item-based top-n recommendation algorithms. In *Proceedings of the Tenth International Conference on Information and Knowledge Management, CIKM '01*, pages 247–254, New York, NY, USA, 2001. ACM. ISBN 1-58113-436-3. doi:10.1145/502585.502627.
- J. Kaur, M. JafariAsbagh, F. Radicchi, and F. Menczer. Scholarometer: A system for crowdsourcing scholarly impact metrics. In *Proceedings of the 2014 ACM Conference on Web Science, WebSci '14*, pages 285–286, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2622-3. doi:10.1145/2615569.2615669.
- R. Kern, C. Korner, and M. Strohmaier. Exploring the influence of tagging motivation on tagging behavior. In M. Lalmas, J. Jose, A. Rauber, F. Sebastiani, and I. Frommholz, editors, *Research and Advanced Technology for Digital Libraries*, volume 6273 of *Lecture Notes in Computer Science*, pages 461–465. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-15463-8. doi:10.1007/978-3-642-15464-5\_57.
- R. Kern, K. Jack, and M. Granitzer. Recommending scientific literature: Comparing use-cases and algorithms. arXiv:1409.1357 [cs.IR], 2014. URL <http://arxiv.org/abs/1409.1357>.
- M. Khabsa and C. L. Giles. The number of scholarly documents on the public web. *PLoS ONE*, 9(5):e93949, May 2014. doi:10.1371/journal.pone.0093949.
- M. Kibanov, M. Atzmueller, C. Scholz, and G. Stumme. On the evolution of contacts and communities in networks of face-to-face proximity. In *Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCoM), IEEE International Conference on and IEEE Cyber, Physical and Social Computing*, pages 993–1000, Aug. 2013. doi:10.1109/GreenCom-iThings-CPSCoM.2013.170.
- M. Kibanov, M. Atzmueller, J. Illig, C. Scholz, A. Barrat, C. Cattuto, and G. Stumme. Is web content a good proxy for real-life interaction? a case study considering online and offline interactions of computer scientists. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 697–704, Boston, MA, USA, 2015. IEEE Press. doi:10.1145/2808797.2810060.

- H.-N. Kim and A. El Saddik. Personalized pagerank vectors for tag recommendations: Inside folkrank. In *Proceedings of the fifth ACM conference on Recommender systems, RecSys '11*, pages 45–52, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0683-6. doi:10.1145/2043932.2043945.
- J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46:604–632, Sept. 1999. doi:10.1145/324133.324140.
- D. Koschützki, K. Lehmann, L. Peeters, S. Richter, D. Tenfelde-Podehl, and O. Zlotowski. Centrality indices. In U. Brandes and T. Erlebach, editors, *Network Analysis*, volume 3418 of *Lecture Notes in Computer Science*, pages 16–61. Springer, Berlin/Heidelberg, 2005. doi:10.1007/978-3-540-31955-9\_3.
- P. Kraker and E. Lex. A critical look at the researchgate score as a measure of scientific reputation. In *Quantifying and Analysing Scholarly Communication on the Web (ASCW'15)*, 2015. doi:10.5281/zenodo.35401.
- B. Krause, C. Schmitz, A. Hotho, and G. Stumme. The anti-social tagger: Detecting spam in social bookmarking systems. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web, AIRWeb '08*, pages 61–68, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-159-0. doi:10.1145/1451983.1451998.
- B. Krause, H. Lerch, A. Hotho, A. Roßnagel, and G. Stumme. Datenschutz im Web 2.0 am Beispiel des sozialen Tagging-Systems BibSonomy. *Informatik-Spektrum*, pages 1–12, 2010. ISSN 0170-6012. doi:10.1007/s00287-010-0485-8.
- R. Krestel, P. Fankhauser, and W. Nejdl. Latent dirichlet allocation for tag recommendation. In *Proceedings of the third ACM conference on Recommender systems*, pages 61–68, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-435-5. doi:10.1145/1639714.1639726.
- M. Kubatz, F. Gedikli, and D. Jannach. Localrank – neighborhood-based, fast computation of tag recommendations. In C. Huemer and T. Setzer, editors, *E-Commerce and Web Technologies*, volume 85 of *Lecture Notes in Business Information Processing*, pages 258–269. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-23013-4. doi:10.1007/978-3-642-23014-1\_22.
- S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 03 1951. doi:10.1214/aoms/1177729694.
- C. Körner, D. Benz, M. Strohmaier, A. Hotho, and G. Stumme. Stop thinking, start tagging: Tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th International World Wide Web Conference (WWW 2010)*, WWW '10, pages 521–530, New York, NY, USA, Apr. 2010. ACM. ISBN 978-1-60558-799-8. doi:10.1145/1772690.1772744.



- 
- E. Lacic, D. Kowald, P. Seitlinger, C. Trattner, and D. Parra. Recommending items in social tagging systems using tag and time informations. In F. Cena, A. S. da Silva, and C. Trattner, editors, *Hypertext 2014 Extended Proceedings Late-breaking Results, Doctoral Consortium and Workshop Proceedings of the 25th ACM Hypertext and Social Media Conference (Hypertext 2014) Santiago, Chile, September 1-4, 2014.*, volume 1210 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014. URL <http://ceur-ws.org/Vol-1210/>.
- A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. *Physical Review E*, 80:056117, Nov 2009. doi:10.1103/PhysRevE.80.056117.
- N. Landia, S. S. Anand, A. Hotho, R. Jäschke, S. Doerfel, and F. Mitzlaff. Extending folkRank with content data. In *Proceedings of the 4th ACM RecSys workshop on Recommender systems and the social web*, RSWeb '12, pages 1–8, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1638-5. doi:10.1145/2365934.2365936.
- D. Lee and P. Brusilovsky. Recommending talks at research conferences using users' social networks. *International Journal of Cooperative Information Systems*, 23(02): 1441003, 2014. doi:10.1142/S0218843014410032.
- D. H. Lee and P. Brusilovsky. Using self-defined group activities for improving recommendations in collaborative tagging systems. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 221–224. ACM, 2010. ISBN 978-1-60558-906-0. doi:10.1145/1864708.1864752.
- J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 915–924, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2. doi:10.1145/1367497.1367620.
- V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707, February 1966. URL [http://profs.sci.univr.it/~liptak/ALBioinfo/2011\\_2012/files/levenshtein66.pdf](http://profs.sci.univr.it/~liptak/ALBioinfo/2011_2012/files/levenshtein66.pdf).
- X. Li, M. Thelwall, and D. Giustini. Validating online reference managers for scholarly impact measurement. *Scientometrics*, 91(2):461–471, 2012. ISSN 0138-9130. doi:10.1007/s11192-011-0580-x.
- J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, Jan. 1991. ISSN 0018-9448. doi:10.1109/18.61115.
- M. Lipczak, Y. Hu, Y. Kollet, and E. Milios. Tag sources for recommendation in collaborative tagging systems. In Eisterlehner et al. [2009], pages 157–172. URL <http://ceur-ws.org/Vol-497>.

- X. Liu, T. Suel, and N. Memon. A robust model for paper reviewer assignment. In *Proceedings of the 8th ACM Conference on Recommender Systems, RecSys '14*, pages 25–32, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2668-1. doi:10.1145/2645710.2645749.
- J. Lorince, K. Joseph, and P. M. Todd. Analysis of music tagging and listening patterns: Do tags really function as retrieval aids? In N. Agarwal, K. Xu, and N. Osgood, editors, *Social Computing, Behavioral-Cultural Modeling, and Prediction*, volume 9021 of *Lecture Notes in Computer Science*, pages 141–152. Springer International Publishing, 2015. ISBN 978-3-319-16267-6. doi:10.1007/978-3-319-16268-3\_15.
- E. D. López-Cózar, N. Robinson-García, and D. Torres-Salinas. Manipulating google scholar citations and google scholar metrics: Simple, easy and tempting. arXiv:1212.0638 [cs.DL], 2012. URL <http://arxiv.org/abs/1212.0638>.
- F. Ma, W. Wang, and Z. Deng. TagRank: A new tag recommendation algorithm and recommender enhancement with data fusion techniques. In S. Zhou and Z. Wu, editors, *Social Media Retrieval and Mining*, volume 387 of *Communications in Computer and Information Science*, pages 80–91. Springer, Berlin/Heidelberg, 2013. ISBN 978-3-642-41628-6. doi:10.1007/978-3-642-41629-3\_7.
- H. Ma, H. Yang, M. R. Lyu, and I. King. SoRec: Social recommendation using probabilistic matrix factorization. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08*, pages 931–940, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi:10.1145/1458082.1458205.
- B. E. Macek, C. Scholz, M. Atzmueller, and G. Stumme. Anatomy of a conference. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media, HT '12*, pages 245–254, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1335-3. doi:10.1145/2309996.2310038.
- M. Manca, L. Boratto, and S. Carta. Friend recommendation in a social bookmarking system: Design and architecture guidelines. In K. Arai, S. Kapoor, and R. Bhatia, editors, *Intelligent Systems in Science and Information 2014*, volume 591 of *Studies in Computational Intelligence*, pages 227–242. Springer International Publishing, 2015. ISBN 978-3-319-14653-9. doi:10.1007/978-3-319-14654-6\_14.
- A. Mandavilli. Peer review: Trial by twitter. *Nautre*, 469(1):286–287, 2011. doi:10.1038/469286a.
- C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, 2008. ISBN 9780521865715 0521865719. URL <http://www.cambridge.org/us/catalogue/catalogue.asp?isbn=9780521865715>.

- 
- B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th International Conference on World Wide Web, WWW '09*, pages 641–641, New York, NY, USA, April 2009. ACM. ISBN 978-1-60558-487-4. doi:10.1145/1526709.1526796.
- C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, HYPERTEXT '06*, pages 31–40, New York, NY, USA, 2006. ACM. ISBN 1-59593-417-0. doi:10.1145/1149941.1149949.
- A. Mas-Bleda, M. Thelwall, K. Kousha, and I. Aguillo. Do highly cited researchers successfully use the social web? *Scientometrics*, 101(1):337–356, 2014. doi:10.1007/s11192-014-1345-0.
- A. Mathes. Folksonomies: Cooperative classification and communication through shared metadata, June 2004. URL <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>. Accessed: 2013-07-11.
- S. M. McNee, N. Kapoor, and J. A. Konstan. Don't look stupid: Avoiding pitfalls when recommending research papers. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work, CSCW '06*, pages 171–180, New York, NY, USA, 2006. ACM. ISBN 1-59593-249-6. doi:10.1145/1180875.1180903.
- P. Melville, R. J. Mooney, and R. Nagarajan. Content-boosted collaborative filtering for improved recommendations. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pages 187–192. AAAI, July 2002. URL <http://www.cs.utexas.edu/users/ai-lab/?melville:aaai02>.
- M. Meriac, A. Fiedler, A. Hohendorf, J. Reinhardt, M. Starostik, and J. Mohnke. Localization techniques for a mobile museum information system. In *Proceedings of WCI (Wireless Communication and Information)*, 2007. URL [http://inka.htw-berlin.de/emika/dokumente/Paper\\_WCI2007.pdf](http://inka.htw-berlin.de/emika/dokumente/Paper_WCI2007.pdf).
- D. R. Millen and J. Feinberg. Using social tagging to improve social navigation. In *Workshop on the Social Navigation and Community based Adaptation Technologies*, 2006. URL [http://www.pitt.edu/~paws/SNC\\_BAT06/crc/millen.pdf](http://www.pitt.edu/~paws/SNC_BAT06/crc/millen.pdf).
- D. R. Millen, M. Yang, S. Whittaker, and J. Feinberg. Social bookmarking and exploratory search. In L. J. Bannon, I. Wagner, C. Gutwin, R. H. Harper, and K. Schmidt, editors, *ECSCW 2007*, pages 21–40. Springer London, 2007. ISBN 978-1-84800-030-8. doi:10.1007/978-1-84800-031-5\_2.
- G. Mishne. Autotag: A collaborative approach to automated tag assignment for weblog posts. In *WWW '06: Proceedings of the 15th International Conference on World Wide Web*, pages 953–954, New York, NY, USA, 2006. ACM Press. ISBN 1595933239. doi:10.1145/1135777.1135961.

- F. Mitzlaff, M. Atzmueller, D. Benz, A. Hotho, and G. Stumme. Community assessment using evidence networks. In M. Atzmueller, A. Hotho, M. Strohmaier, and A. Chin, editors, *Analysis of Social Media and Ubiquitous Data*, volume 6904 of *Lecture Notes in Computer Science*, pages 79–98. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-23598-6. doi:10.1007/978-3-642-23599-3\_5.
- E. Mohammadi and M. Thelwall. Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. *Journal of the Association for Information Science and Technology*, 65(8):1627–1638, 2014. ISSN 2330-1643. doi:10.1002/asi.23071.
- E. Mohammadi, M. Thelwall, S. Haustein, and V. Larivière. Who reads research articles? an altmetrics analysis of Mendeley user categories. *Journal of the Association for Information Science and Technology*, 66(9):1832–1846, 2015. ISSN 2330-1643. doi:10.1002/asi.23286.
- E. Montañés, J. Ramón Quevedo, I. Díaz, R. Cortina, P. Alonso, and J. Ranilla. TagRanker: Learning to recommend ranked tags. *Logic Journal of IGPL*, 19(2):395–404, 2011. doi:10.1093/jigpal/jzq036.
- S. Mossa, M. Barthélémy, H. E. Stanley, and L. A. N. Amaral. Truncation of power law behavior in “scale-free” network models due to information filtering. *Physical Review Letters*, 88(13):138701, Mar 2002. doi:10.1103/PhysRevLett.88.138701.
- C. Musto, F. Narducci, P. Lops, and M. de Gemmis. Combining collaborative and content-based techniques for tag recommendation. In *E-Commerce and Web Technologies*, volume 61 of *Lecture Notes in Business Information Processing*, pages 13–23, Berlin/Heidelberg, 2010. Springer. ISBN 978-3-642-15207-8. doi:10.1007/978-3-642-15208-5\_2.
- U. T. Müller. *Peer-Review-Verfahren zur Qualitätssicherung von Open-Access-Zeitschriften – Systematische Klassifikation und empirische Untersuchung*. PhD thesis, Humboldt-Universität zu Berlin, 2008. URL <http://edoc.hu-berlin.de/dissertationen/mueller-uwe-thomas-2008-12-17/PDF/mueller.pdf>.
- M. Nagelschmidt. Literaturverwaltungsprogramme – Das funktionale Minimum. *ABI-Technik*, 30(2):94–99, June 2010. ISSN 21914664. doi:10.1515/ABITECH.2010.30.2.94.
- A. Nanopoulos, D. Rafailidis, and I. Karydis. Matrix factorization with content relationships for media personalization. In *Wirtschaftsinformatik Proceedings 2013*, pages 87–101, Mar. 2013. URL <http://aisel.aisnet.org/wi2013/6/>.
- M. E. J. Newman. Analysis of weighted networks. *PHYSICAL REVIEW E*, 70(5):056131, Nov. 2004. doi:10.1103/PhysRevE.70.056131.

- 
- M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006. doi:10.1073/pnas.0601602103.
- M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, Feb. 2004. doi:10.1103/PhysRevE.69.026113.
- D. Nicholas, P. Huntington, and A. Watkinson. Scholarly journal usage: The results of deep log analysis. *Journal of Documentation*, 61(2):248–280, 2005. doi:10.1108/00220410510585214.
- N. F. Noy, A. Chugh, and H. Alani. The ckc challenge: Exploring tools for collaborative knowledge construction. *IEEE Intell Syst*, 23(1):64–68, 1 2008. doi:10.1109/MIS.2008.14.
- E. Orduna-Malea, J. M. Ayllon, A. Martin-Martin, and E. D. Lopez-Cozar. Empirical evidences in citation-based search engines: Is Microsoft Academic Search dead? arXiv:1404.7045 [cs.DL], 2014. URL <http://arxiv.org/abs/1404.7045>.
- J. L. Ortega. Relationship between altmetric and bibliometric indicators across academic social sites: The case of CSIC’s members. *Journal of Informetrics*, 9(1): 39–49, 2015. doi:10.1016/j.joi.2014.11.004.
- D. Parra and P. Brusilovsky. Evaluation of collaborative filtering algorithms for recommending articles on CiteULike. In *Proceedings of the Workshop on Web 3.0: Merging Semantic Web and Social Web*, volume 467 of *CEUR Workshop Proceedings*, 2009. URL <http://ceur-ws.org/Vol-467/paper5.pdf>.
- D. Parra, W. Jeng, P. Brusilovsky, C. López, and S. Sahebi. Conference navigator 3: An online social conference support system. In *Workshop and Poster Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization Montreal, Canada, July 16-20, 2012*, volume 872 of *CEUR Workshop Proceedings*, 2012. URL <http://ceur-ws.org/Vol-872/>.
- N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1):25–46, 1999. ISSN 0306-4379. doi:10.1016/S0306-4379(99)00003-4.
- I. Peters. *Folksonomies. Indexing and Retrieval in Web 2.0*. Knowledge & Information - Studies in Information Theory. De Gruyter/Saur, Berlin, 2009. ISBN 9783598441851. doi:10.1515/9783598441851.
- I. Peters and W. G. Stock. “power tags” in information retrieval. *Library Hi Tech*, 28 (1):81–93, 2010. doi:10.1108/07378831011026706.
- I. Peters and K. Weller. Tag gardening for folksonomy enrichment and maintenance. *Webology*, 5(3), 2008. URL <http://www.webology.org/2008/v5n3/a58.html>.

- I. Peters, S. Haustein, and J. Terliesner. Crowdsourcing in article evaluation. In *3rd International Conference on Web Science (ACM WebSci'11)*, pages 1–4, June 2011. URL [http://www.websci11.org/www.websci11.org/fileadmin/websci/Posters/89\\_paper.pdf](http://www.websci11.org/www.websci11.org/fileadmin/websci/Posters/89_paper.pdf).
- I. Peters, L. Schumann, and J. Terliesner. Folksonomy-basiertes Information Retrieval unter der Lupe. *Information – Wissenschaft & Praxis*, 63:273–280, 2012. ISSN 16194292. doi:10.1515/iwp-2012-0047.
- I. Peters, P. Kraker, E. Lex, C. Gumpenberger, and J. Gorraiz. Research data explored: Citations versus altmetrics. In *Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 29 June to 3 July, 2015*, pages 172–183. Bogaziçi University Printhouse, 2015. URL <http://www.issi2015.org/files/downloads/all-papers/0172.pdf>.
- W. Petersen and P. Heinrich. Qualitative citation analysis based on formal concept analysis. Presented at the 32th annual meeting of the Classification Society in Hamburg, 07 2008. URL [http://user.phil-fak.uni-duesseldorf.de/~petersen/slides/Petersen\\_Heinrich\\_GFKL2008\\_slides.pdf](http://user.phil-fak.uni-duesseldorf.de/~petersen/slides/Petersen_Heinrich_GFKL2008_slides.pdf).
- T. J. Phelan. A compendium of issues for citation analysis. *Scientometrics*, 45(1): 117–136, 1999. ISSN 0138-9130. doi:10.1007/BF02458472.
- J. Poelmans, P. Elzinga, S. Viaene, and G. Dedene. Formal concept analysis in knowledge discovery: A survey. In M. Croitoru, S. Ferré, and D. Lukose, editors, *Conceptual Structures: From Information to Intelligence*, volume 6208 of *Lecture Notes in Computer Science*, pages 139–153. Springer, Berlin/Heidelberg, 2010. doi:10.1007/978-3-642-14197-3\_15.
- J. Poelmans, D. I. Ignatov, S. Viaene, G. Dedene, and S. O. Kuznetsov. Text mining scientific papers: A survey on fca-based information retrieval research. In P. Perner, editor, *Advances in Data Mining. Applications and Theoretical Aspects: 12th Industrial Conference, ICDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings*, pages 273–287, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-31488-9. doi:10.1007/978-3-642-31488-9\_22.
- J. Poelmans, D. I. Ignatov, S. O. Kuznetsov, and G. Dedene. Formal concept analysis in knowledge processing: A survey on applications. *Expert Systems with Applications*, 40(16):6538–6560, 2013a. ISSN 0957-4174. doi:10.1016/j.eswa.2013.05.009.
- J. Poelmans, S. O. Kuznetsov, D. I. Ignatov, and G. Dedene. Formal concept analysis in knowledge processing: A survey on models and techniques. *Expert Systems with Applications*, 40(16):6601–6623, 2013b. ISSN 0957-4174. doi:10.1016/j.eswa.2013.05.007.
- S. Pohl, F. Radlinski, and T. Joachims. Recommending related papers based on digital library access records. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference*

- 
- on *Digital Libraries*, JCDL '07, pages 417–418, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-644-8. doi:10.1145/1255175.1255260.
- A. Popescul, L. H. Ungar, D. M. Pennock, and S. Lawrence. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 437–444, San Francisco, CA, USA, Aug. 2001. Morgan Kaufmann Publishers Inc. URL [http://repository.upenn.edu/cgi/viewcontent.cgi?article=1144&context=cis\\_papers](http://repository.upenn.edu/cgi/viewcontent.cgi?article=1144&context=cis_papers).
- J. Porter. Learning more about structured blogging, 2005. URL <http://bokardo.com/archives/learning-more-about-structured-blogging/>. Accessed: 2013-08-12.
- M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980. doi:10.1108/eb046814.
- Pressestelle des Bundesgerichtshofs. Bundesgerichtshof entscheidet über die Zulässigkeit einer Lehrerbewertung im Internet ([www.spickmich.de](http://www.spickmich.de)). *Bundesgerichtshof – Mitteilung der Pressestelle*, 2009(137), 2009. URL <http://juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=pm&Datum=2009&Sort=3&nr=48373>.
- Pressestelle des Bundesgerichtshofs. Bundesgerichtshof lehnt den Anspruch eines Arztes auf Löschung seiner Daten aus einem Ärztebewertungsportal ab. *Bundesgerichtshof – Mitteilung der Pressestelle*, 2014(132), 2014. URL <http://juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=pm&Datum=2014&Sort=3&nr=68915&anz=132>.
- D. J. d. S. Price. *Little science, big science*. Columbia University Press, New York, 1963. ISBN 0231085621 9780231085625. URL <http://www.worldcat.org/oclc/522357>.
- J. Priem and B. H. Hemminger. Scientometrics 2.0: New metrics of scholarly impact on the social web. *First Monday*, 15(7), July 2010. URL <http://journals.uic.edu/ojs/index.php/fm/article/view/2874/2570>.
- J. Priem, D. Taraborelli, P. Groth, and C. Neylon. Altmetrics: A manifesto, 2010. URL <http://altmetrics.org/manifesto/>.
- J. Priem, H. A. Piwowar, and B. M. Hemminger. Altmetrics in the wild: Using social media to explore scholarly impact. arXiv:1203.4745 [cs.DL], 2012. URL <http://arxiv.org/abs/1203.4745>.
- M. Ramezani. Improving graph-based approaches for personalized tag recommendation. *Journal of Emerging Technologies in Web Intelligence*, 3(2):168–176, 2011. doi:10.4304/jetwi.3.2.168-176.

- M. Ramezani, J. Gemmell, T. Schimoler, and B. Mobasher. Improving link analysis for tag recommendation in folksonomies. In *2nd ACM RecSys10 Workshop on Recommender Systems and the Social Web, Barcelona, 2010*, 2010. URL <http://josquin.cs.depaul.edu/~mramezani/papers/2010recsys.pdf>.
- S. Redner. How popular is your paper? an empirical study of the citation distribution. *European Physical Journal B*, 4(2):131–134, Aug. 1998. doi:10.1007/s100510050359.
- S. Rendle and L. Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM '10*, pages 81–90. ACM, 2010. ISBN 978-1-60558-889-6. doi:10.1145/1718487.1718498.
- S. Rendle, L. Balby Marinho, A. Nanopoulos, and L. Schmidt-Thieme. Learning optimal ranking with tensor factorization for tag recommendation. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09*, pages 727–736, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-495-9. doi:10.1145/1557019.1557100.
- T. Rock and R. Wille. Ein TOSCANA-Erkundungssystem zur Literatursuche. FB4-Preprint 1901, TH Darmstadt, 1997. URL <http://www3.mathematik.tu-darmstadt.de/fb/mathe/preprints.html>.
- M. Rosvall and C. T. Bergstrom. An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104(18):7327–7331, 2007. doi:10.1073/pnas.0611034104.
- L. Sachs. *Statistische Methoden: Planung und Auswertung*. Springer-Verlag, Berlin, 1988. ISBN 038718113X 9780387181134 354018113X 9783540181132. URL <http://www.worldcat.org/oclc/613645241>.
- A. Saeed, M. Afzal, A. Latif, and K. Tochtermann. Citation rank prediction based on bookmark counts: Exploratory case study of WWW06 papers. In *Multitopic Conference, 2008. INMIC 2008. IEEE International*, pages 392–397, Dec. 2008. doi:10.1109/INMIC.2008.4777769.
- A. Said and A. Bellogín. Comparative recommender system evaluation: Benchmarking recommendation frameworks. In *Proceedings of the 8th ACM Conference on Recommender Systems, RecSys '14*, pages 129–136, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2668-1. doi:10.1145/2645710.2645746.
- S. L. Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1(3):317–328, 1997. ISSN 1384-5810. doi:10.1023/A:1009752403260.



- 
- B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW '01 Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM, 2001. ISBN 1-58113-348-0. doi:10.1145/371920.372071.
- B. M. Sarwar, G. Karypis, J. A. Konstan, and J. T. Riedl. Application of dimensionality reduction in recommender system – a case study. In *ACM WEBKDD Workshop*, 2000. URL <http://files.grouplens.org/papers/webKDD00.pdf>.
- A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '02*, pages 253–260, New York, NY, USA, 2002. ACM. ISBN 1-58113-561-0. doi:10.1145/564376.564421.
- F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger. Understanding online social network usage from a network perspective. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference, IMC '09*, pages 35–48, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-771-4. doi:10.1145/1644893.1644899.
- C. Scholz, S. Doerfel, M. Atzmueller, A. Hotho, and G. Stumme. Resource-aware on-line rfid localization using proximity data. In *ECML/PKDD (3)*, pages 129–144, 2011. doi:10.1007/978-3-642-23808-6\_9.
- C. Scholz, M. Atzmueller, A. Barrat, C. Cattuto, and G. Stumme. New insights and methods for predicting face-to-face contacts. In *Seventh International AAAI Conference on Weblogs and Social Media*, Palo Alto, CA, USA, 2013. AAAI Press. URL <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6097/6396>.
- C. Scholz, J. Illig, M. Atzmueller, and G. Stumme. On the predictability of talk attendance at academic conferences. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media, HT '14*, pages 279–284, New York, NY, USA, 2014. ACM Press. ISBN 978-1-4503-2954-5. doi:10.1145/2631775.2631816.
- C. Schuh. *Publikationsverhalten im Überblick – eine Zusammenfassung der einzelnen Diskussionsbeiträge*, pages 6–13. Diskussionspapiere der Alexander von Humboldt-Stiftung. Alexander von Humboldt-Stiftung, Bonn, 2 edition, Dec 2009. URL [https://www.humboldt-foundation.de/pls/web/docs/F13905/12\\_disk\\_papier\\_publicationsverhalten2\\_kompr.pdf](https://www.humboldt-foundation.de/pls/web/docs/F13905/12_disk_papier_publicationsverhalten2_kompr.pdf).
- J. Scripps, P.-N. Tan, and A.-H. Esfahanian. Exploration of link structure and community-based node roles in network analysis. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 649–654, Oct 2007a. doi:10.1109/ICDM.2007.37.

- J. Scripps, P.-N. Tan, and A.-H. Esfahanian. Node roles and community structure in networks. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 26–35, New York, NY, USA, 2007b. ACM. ISBN 978-1-59593-848-0. doi:10.1145/1348549.1348553.
- S. B. Seidman. Network structure and minimum degree. *Social Networks*, 5(3):269–287, 1983. ISSN 0378-8733. doi:10.1016/0378-8733(83)90028-X.
- P. Seitlinger, D. Kowald, C. Trattner, and T. Ley. Recommending tags with a model of human categorization. In *Proceedings of the 22nd International Conference on Conference on Information & Knowledge Management, CIKM '13*, pages 2381–2386, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2263-8. doi:10.1145/2505515.2505625.
- G. Shani and A. Gunawardana. Evaluating recommendation systems. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 257–297. Springer US, 2011. doi:10.1007/978-0-387-85820-3\_8.
- A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM conference on Recommender systems, RecSys '08*, pages 259–266. ACM, 2008. ISBN 978-1-60558-093-7. doi:10.1145/1454008.1454048.
- J. Sinclair and M. Cardew-Hall. The folksonomy tag cloud: When is it useful? *Journal of Information Science*, 34(1):15–29, 2008. doi:10.1177/0165551506078083.
- S. Sood, S. Owsley, K. Hammond, and L. Birnbaum. TagAssist: Automatic tag suggestion for blog posts. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, Boulder, Colorado, USA, 2007. URL <http://www.icwsml.org/papers/2--Sood-Owsley-Hammond-Birnbaum.pdf>.
- J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, and P. Vanhems. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE*, 6(8):1–13, 08 2011. doi:10.1371/journal.pone.0023176.
- J. Stiller, M. Gäde, and V. Petras. Is tagging multilingual?: A case study with BibSonomy. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, JCDL '11*, pages 421–422, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0744-4. doi:10.1145/1998076.1998165.
- M. Strohmaier, C. Körner, and R. Kern. Why do users tag? detecting users' motivation for tagging in social tagging systems. In *Fourth International AAAI Conference on Weblogs and Social Media, ICWSM '10*, 2010. URL <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1497>.

- 
- G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhali. Computing iceberg concept lattices with titanic. *Data & Knowledge Engineering*, 42(2):189–222, Aug. 2002. doi:10.1016/S0169-023X(02)00057-5.
- R. Tagliacozzo. Self-citations in scientific literature. *Journal of Documentation*, 33(4): 251–265, 1977. doi:10.1108/eb026644.
- J. Tague-Sutcliffe. An introduction to informetrics. *Information Processing & Management*, 28(1):1–3, 1992. ISSN 0306-4573. doi:10.1016/0306-4573(92)90087-G.
- D. Taraborelli. Soft peer review: Social software and distributed scientific evaluation. In *Proceedings of the 8th International Conference on the Design of Cooperative Systems (COOP '08)*, 2008. URL [http://coop.wineme.fb5.uni-siegen.de/proceedings2008/3\\_02\\_dTaraborelli\\_al\\_99-110.pdf](http://coop.wineme.fb5.uni-siegen.de/proceedings2008/3_02_dTaraborelli_al_99-110.pdf).
- D. Terdiman. Folksonomies tap people power, Jan. 2005. URL <http://www.wired.com/2005/02/folksonomies-tap-people-power>. Accessed: 2013-08-12.
- M. Thelwall. Journal impact evaluation: A webometric perspective. *Scientometrics*, pages 1–13, 2012. ISSN 0138-9130. doi:10.1007/s11192-012-0669-x.
- M. Thelwall. Why do papers have many Mendeley readers but few Scopus-indexed citations and vice versa? *Journal of Librarianship and Information Science*, 2015. doi:10.1177/0961000615594867.
- M. Thelwall and R. Fairclough. The influence of time and discipline on the magnitude of correlations between citation counts and quality scores. *Journal of Informetrics*, 9(3):529–541, 2015. doi:10.1016/j.joi.2015.05.006.
- M. Thelwall and P. Sud. Mendeley readership counts: An investigation of temporal and disciplinary differences. *Journal of the Association for Information Science and Technology*, pages 3036–3050, 2015. ISSN 2330-1643. doi:10.1002/asi.23559.
- M. Thelwall and P. Wilson. Distributions for cited articles from individual subjects and years. *Journal of Informetrics*, 8(4):824–839, 2014. ISSN 1751-1577. doi:10.1016/j.joi.2014.08.001.
- M. Thelwall and P. Wilson. Mendeley readership altmetrics for medical articles: An analysis of 45 fields. *Journal of the Association for Information Science and Technology*, pages 1962–1972, 2015. ISSN 2330-1643. doi:10.1002/asi.23501.
- M. Thelwall, S. Haustein, V. Larivière, and C. R. Sugimoto. Do altmetrics work? Twitter and ten other social web services. *PLoS ONE*, 8(5):e64841, 05 2013. doi:10.1371/journal.pone.0064841.
- B. Thijs and W. Glänzel. The influence of author self-citations on bibliometric meso-indicators. the case of european universities. *Scientometrics*, 66(1):71–80, 2006. ISSN 1588-2861. doi:10.1007/s11192-006-0006-3.

- P. Thomas. Using interaction data to explain difficulty navigating online. *ACM Transactions on the Web*, 8(4):24:1–24:41, Nov. 2014. ISSN 1559-1131. doi:10.1145/2656343.
- T. Tilley, R. Cole, P. Becker, and P. Eklund. A survey of formal concept analysis support for software engineering activities. In B. Ganter, G. Stumme, and R. Wille, editors, *Formal Concept Analysis*, pages 250–271. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 3-540-27891-5, 978-3-540-27891-7. doi:10.1007/11528784\_13.
- R. Torres, S. McNee, M. Abel, J. Konstan, and J. Riedl. Enhancing digital libraries with techlens. In *Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference on*, pages 228–236, June 2004. doi:10.1109/JCDL.2004.1336126.
- J. Trant. Studying social tagging and folksonomy: A review and framework. *Journal of Digital Information*, 10(1), 2009. ISSN 1368-7506. URL <https://journals.tdl.org/jodi/index.php/jodi/article/view/269>.
- T. Vander Wal. Tagging for fun and finding, July 2005. URL <http://okcancel.com/archives/article/2005/07/tagging-for-fun-and-finding.html>. Accessed: 2013-08-12.
- T. Vander Wal. Folksonomy, 2007. URL <http://vanderwal.net/folksonomy.html>. Accessed: 2013-08-12.
- M. Vojnović, J. Cruise, D. Gunawardena, and P. Marbach. Ranking and suggesting tags in collaborative tagging applications. Technical Report MSR-TR-2007-06, Microsoft Research, Feb. 2007. URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.6786&rep=rep1&type=pdf>.
- J. Voss, A. Hotho, and R. Jäschke. Mapping bibliographic records with bibliographic hash keys. In R. Kuhlen, editor, *Information: Droge, Ware oder Commons?*, Proceedings of the ISI. Hochschulverband Informationswissenschaft, Verlag Werner Hülsbusch, 2009. URL <http://hdl.handle.net/10760/12697>.
- Q. H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333, 1989. ISSN 00129682. doi:10.2307/1912557.
- L. Waltman and R. Costas. F1000 recommendations as a potential new data source for research evaluation: A comparison with citations. *Journal of the Association for Information Science and Technology*, 65(3):433–445, 2014. ISSN 2330-1643. doi:10.1002/asi.23040.
- J.-C. Wang and C.-C. Chiu. Recommending trusted online auction sellers using social network analysis. *Expert Systems with Applications*, 34(3):1666–1679, 2008. ISSN 0957-4174. doi:10.1016/j.eswa.2007.01.045.
- C. Wartena and M. Wibbels. Improving tag-based recommendation by topic diversification. In *Advances in Information Retrieval*, volume 6611 of *LNCIS*, pages 43–54.

- 
- Springer, Berlin/Heidelberg, 2011. ISBN 978-3-642-20160-8. doi:10.1007/978-3-642-20161-5-7.
- D. Weinberger. Tagging and why it matters. *SSRN eLibrary*, 2005. doi:10.2139/ssrn.870594.
- C. Weller. Beobachtungen wissenschaftlicher Selbstkontrolle. Qualität, Schwächen und die Zukunft des Peer Review-Verfahrens. *Zeitschrift für Internationale Beziehungen*, 11(2):365–394, 2004. ISSN 09467165. URL <http://www.jstor.org/stable/40843977>.
- R. Wetzker, W. Umbrath, and A. Said. A hybrid approach to item recommendation in folksonomies. In *Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '09, pages 25–29, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-430-0. doi:10.1145/1506250.1506255.
- A. Whitby, A. Jøsang, and J. Indulska. Filtering out unfair ratings in bayesian reputation systems. In *Proceedings of the Workshop on Trust in Agent Societies, at the Autonomous Agents and Multi Agent Systems Conference (AAMAS2004)*, 2004. URL <http://folk.uio.no/josang/papers/WJI2004-AAMAS.pdf>.
- R. Wille. Restructuring lattice theory: An approach based on hierarchies of concepts. In I. Rival, editor, *Ordered Sets*, volume 83 of *NATO Advanced Study Institutes Series*, pages 445–470. Springer Netherlands, 1982. ISBN 978-94-009-7800-3. doi:10.1007/978-94-009-7798-3\_15.
- C. Wongchokprasitti, P. Brusilovsky, and D. Para. Conference navigator 2.0: Community-based recommendation for academic conferences. In *Workshop Social Recommender Systems, SRS'10, 7 February 2010, Hong Kong, China*, 2010. URL <http://www.comp.hkbu.edu.hk/~lichen/srs2010/downloads/paper/21-final%20version%20cn20.pdf>.
- P. Wouters and R. Costas. Users, narcissism and control – tracking the impact of scholarly publications in the 21st century. SURF-foundation, 2012. URL <http://research-acumen.eu/wp-content/uploads/Users-narcissism-and-control.pdf>.
- T. Wray and P. Eklund. Exploring the information space of cultural collections using formal concept analysis. In P. Valtchev and R. Jäschke, editors, *Formal Concept Analysis*, volume 6628 of *Lecture Notes in Computer Science*, pages 251–266. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-20513-2. doi:10.1007/978-3-642-20514-9\_19.
- Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. In *Proceedings of the Collaborative Web Tagging Workshop at the WWW 2006*, Edinburgh, Scotland, 2006. URL <http://ra.ethz.ch/CDstore/www2006/www.rawsugar.com/www2006/13.pdf>.

- Z. Zahedi, R. Costas, and P. Wouters. Do Mendeley readership counts help to filter highly cited WoS publications better than average citation impact of journals (JCS)? In *Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 29 June to 3 July, 2015*, pages 16–25. Bogaziçi University Printhouse, 2015. URL <http://www.issi2015.org/files/downloads/all-papers/0016.pdf>.
- H. Zhang, H. Zhao, W. Cai, J. Liu, and W. Zhou. Using the k-core decomposition to analyze the static structure of large-scale software systems. *The Journal of Supercomputing*, 53(2):352–369, 2010. ISSN 0920-8542. doi:10.1007/s11227-009-0299-0.
- D. Zoller, S. Doerfel, R. Jäschke, G. Stumme, and A. Hotho. On publication usage in a social bookmarking system. In *Proceedings of the ACM Web Science Conference, WebSci '15*, pages 67:1–67:2, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3672-7. doi:10.1145/2786451.2786927.
- D. Zoller, S. Doerfel, R. Jäschke, G. Stumme, and A. Hotho. Posted, visited, exported: Altmetrics in the social tagging system BibSonomy. *Journal of Informetrics*, 10(3): 732–749, 2016. ISSN 1751-1577. doi:10.1016/j.joi.2016.03.005.