Enhancing Interdisciplinary Cooperation by Social Platforms

Assessing the Usefulness of Bibliometric Social Network Visualization in Large-Scale Research Clusters

André Calero Valdez[1], Anne Kathrin Schaar[1], Martina Ziefle[1], and Andreas Holzinger[2]

[1] Human-Computer Interaction Center, Campus Boulevard 57, RWTH Aachen University, Germany
{calero-valdez,schaar,ziefle}@comm.rwth-aachen.de
[2] Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Austria
andreas.holzinger@medunigraz.at

Abstract. In large-scale research projects active management of the cooperation process is necessary, e.g. to ensure systematic transfer of knowledge, alignment of research goals, or appropriate dissemination of research efforts. In a large scale research-cluster at the RWTH Aachen University a cybernetic management approach is applied. As a planned measure, publishing efforts (i.e. bibliometric data) will be visualized on a social software platform accessible by researchers and the steering committee. But do researchers agree with the chosen style of visualization of their publications? As part of a user centered design, this paper presents the results of an interview study with researchers (n=22) addressing the usefulness and applicability of this approach. As central findings arguments for using the publication visualization are identified such as enabling *retrospective analysis*, acquiring *new information* about the team, improvement in dissemination *planning*, but at the same time contrasted by arguments against this approach, such as *missing information*, a possibly *negative influence on workflow* of researchers, and the *bad legibility* of the visualization. Additionally requirements and suggested improvements are presented.

Keywords: Data visualization, technology acceptance, bibliometrics, user centred-design, information systems.

1 Introduction and Motivation for Research

Large-scale research problems like health and aging or economics and production in high wage countries are no longer solvable by single disciplines or subject areas. Confronted with this, the trend to interdisciplinary research compounds gained more and more influence [1]. Interdisciplinary cooperation is perceived

S. Yamamoto (Ed.): HIMI 2014, Part I, LNCS 8521, pp. 298–309, 2014.

to bring along a multitude of knowhow and more innovative power than disciplinary cooperation. But mere interdisciplinarity is no guaranteed success factor. Bringing together different disciplines often brings along challenges that can burden, disturb or even scupper these actions [2]. Due to the fact that interdisciplinary research projects are promoted systematically by research funding agencies, universities, and industry, it is essential to understand the sensitive points of interdisciplinary cooperation and find adequate measures that support involved researchers, reviewers and management/supervisors to overcome theses challenges by offering guidelines, tools, and rules. The complexity and social nature of these challenges call for decentralized means of communication like a social portal.

In this paper we present an approach to use a social portal software to enhance the interdisciplinary cooperation in a research cluster (Cluster of Excellence "Integrative Production Technologies for High-Wage Countries" at the RWTH Aachen University in Germany). The presented portal is a project of the so-called Cross-Sectional Processes within this research cluster that were implemented to support "networking processes and strategic cluster developments by means of learning and knowledge management" [3]. The central goal of the portal is to support researchers in interdisciplinary large-scale research projects by addressing (among others) three aspects of interdisciplinary challenges. One of them are different uses of terminology, which are supported by using an online project specific glossary on the portal. Another part is a technology portal [4], which allows members of the cluster to exchange key parameters of developed technologies between projects. The thirds is a publication visualization tool, which can be used to understand changes in social structure indicated by publication behavior or other sociometric data. Additionally to portal offers typical social features such as member yellow pages, news feeds, topic based groups, and many more. Since all these features are interconnected (e.g. users can click on terminology entries and find the creators of the entries), it is important to evaluate each tool in context with the whole platform.

In this paper we present a study on the third feature – the publication visualization tool. Using an interview approach we tried to understand what concerns users might have regarding such a visualization and what benefits they would see with it. Additionally we wanted to find out whether the future users had ideas for improvement.

2 Related Work

In the context of the presented study general aspects from the field of bibliometrics and scientometrics must be considered. Therefore this section presents two approaches used for publication analysis (list-based and mapping-based approaches). Lastly the applied publications-visualization approach for this research is presented. This approach was developed in the context of the research cluster Integrative Production Technologies for High-wage Countries to support interdisciplinary work.

List-Based Bibliometrics. Making bibliographic efforts visible to allow researchers to understand their publication behavior has been approached by many using different approaches. Generally two types of strategies can be discerned. List-based analyses created by databases like Google Scholar, Web of Science, or Scopus give the users insight into their citation records, and how well their work is being cited. Results are presented as lists, hence the name, which can be ordered according to criteria like *most cited*, *most recent*, etc. This list-based approach has been used and constantly debated over sixty years[5]. Database approaches always bring along the problem of *database coverage* [6]. In order to track citation records accurately the database providers need to scan millions of documents, identify citations and assign them to individual papers. As a researcher interested in their own citation records must pick a database, that covers the relevant publications and that are likely to cite ones articles. Different citations indices include different types of documents or outlets and might differ in their accuracy. First technical difficulties exist. Identifying a citation correctly from PDF-data, mapping it to a unique record, and finding unique authors is computationally hard. Authors may vary their citation style, make errors in their bibliographies and many researchers have similar or equal names. All this leads to differing levels of coverage [7]. Coverage ratios also depend on the discipline of the authors [8] as disciplines vary in their publishing behavior.

Even when "correct" citation records exist, it is hard to understand what they mean, when trying to relate them to a researchers performance. Disciplines are different according to sheer size, citations per paper, citation half-life and other aspects that require a normalization process to make citations between disciplines comparable [9]. But what is a discipline? The question of *subject delineation* can either be solved by assigning certain outlets to disciplines (e.g. according to their description) or to perform citation analysis, to find coherent structures of citation networks, that are then considered disciplines. Beyond the technical difficulties manipulation (e.g. self-citations, exploiting the algorithm) of data can become a problem for some databases [10]. Even beyond these difficulties, citations can both indicate agreement or disagreement. Sometimes even honorable mentions exist, without adding to the content. Thus extracting the "meaning" of a citations is also computationally hard (i.e. sentiment analysis).

Mapping-Based Bibliometrics. Mapping-based approaches [11] try to visualize publication data in graphs in order to understand the data both visually but also mathematically from a different perspectives. Mapping can be achieved by mapping citations (i.e. who cites who), co-authorship (i.e. who writes with whom), co-citations (i.e. who gets mutually cited in a document) and many more. The relationship is assigned an edge that connected two vertices that represent the item under analysis. Using this approach allows different forms of analyses that are graph based (e.g. centrality, entropy [12], etc.). Nonetheless mapping-based approaches may also suffer from the same problems as list based approaches (i.e database coverage, sentiment analysis).

The Approach of Mixed Node Publicaion Analysis. For the reasons given above, our approach focuses on a mapping based approach and data collection is done manually by the university library (i.e. researchers are required to report their publications). We decided to not visualize citations, but focus on cooperation based measures such as co-authorship, since interpretation of citations was out of scope for our approach. The approach of our visualization tool was designed to make interdisciplinary work more visible, analyzable and steerable. To realize this, a so-called mixed-node graph visualization [13] was conceived, which allows visualizing authors, their publications, as well as their discipline (see Fig. 1) in a single graph. As Fig. 1 illustrates publications are depicted by little, authors by medium sized and the authors' disciplines by large nodes.

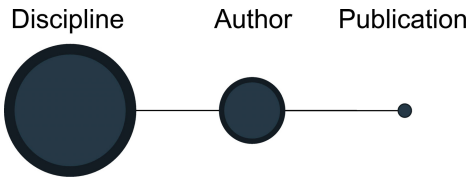

Fig. 1. Mixed node publication graph with different types of nodes. Source: [13]

When coloring the nodes that represent the authors according to their discipline, and running a force based layout algorithm (e.g. Force Atlas 2 from the Gephi Suite [14]), connectedness to other researchers and their disciplines can be analyzed visually (see Fig. 2 and see also `http://vimeo.com/48446978`).

Knowing that user acceptance is the essential key for the success of technological applications we have run (N=22) semi-structured interviews to find out what potential users (i.e. interdisciplinary researchers) thought about the visualization approach, continuing previous research efforts [15].

3 Methodology

We conducted twenty two interviews with researchers from two projects at the RWTH Aachen University. Interviews were divided into four main parts. Part one contained questions about the validity of the tool to accurately represent interdisciplinary team performance. Part two addressed the suitability of the approach to be used as a steering instrument for interdisciplinary research groups. Part three asked for an evaluation of the impact (positive vs. negative) of our approach on the work climate in interdisciplinary work. Part four focused on the evaluation of the approach as a tool for self-measurement for researchers, to locate themselves within their team, as well as to analyze their performance or search for cooperation partners.

The introduction of the interview was a short presentation of a prototype of the visualization tool. The presented visualization was a depiction of the publications of the sample team of the interviewee. The presented visualization showed

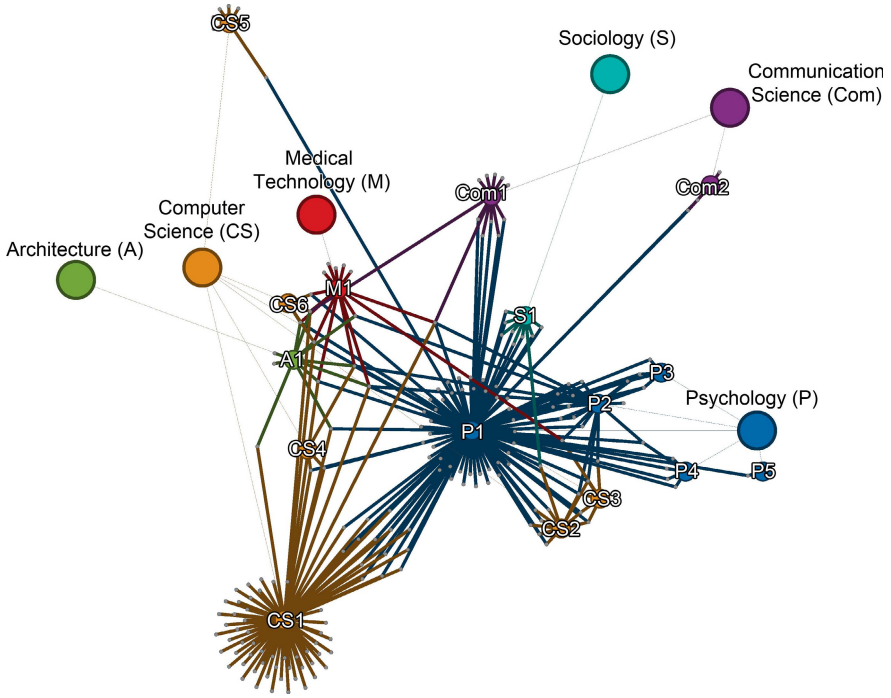

Fig. 2. Sample graph for a workgroup. Researchers are anonymized and represented as numbers according to their discipline. Source: [13]

all names of the team members as well as their disciplines and publication titles (due to privacy see an anonymized example in Fig. 2), generated according to the reduced graph described by Calero Valdez et al. [13]. After the presentation of the prototype, participants were asked to evaluate the prototype of the visualization tool.

The interviews were recorded and transcribed and then scanned for open codes. The open codes were then conceptualized and summarized into categories [16]. For each category, numbers of mentions were counted to establish relative importance. These steps were performed for both arguments for and against the tool. Additionally requirements and wishes for the tool were collected as open codes.

4 Results

The analysis of the transcript lead to the identification of 26 underlying concepts that contained arguments for and against the use of the tool (see Table 1). After identifying these concepts numbers of mentions were counted in the transcript, yielding a total of 139 mentions of the concepts. In total 76 mentions were counted for pro arguments against 63 mentions for con arguments.

Table 1. Examplary (translated) transcripts and the mapping to concepts

Concept	Transcript text
Pro 1) Retrospective analysis	*"One can say he did everything he should have done, if you look at it divided over the years."*
Pro 2) Information regarding the team	*"[...] when I see myself related to the others, it mirrors the degree of cooperation."*
Pro 3) Planning	*"It is interesting to look at for yourself and the head of the institute and to find blind spots and develop or strengthen relationships."*
Con 1) Missing information	*"There is no information about the impact factor or who the first author is."*
Con 2) Negative influence on workflow	*"At last it might downgrade all the colleagues to little atoms which wander around the two big atoms."*
Con 3) Bad legibility	*"It's hard to see with whom you've published, because there are so many lines."*

4.1 Pro Arguments

Our results revealed that interdisciplinary working researchers have a generally positive attitude towards our visualizations approach. Main benefits are seen in the chance to analyze the group and own work retrospectively (15 mentions), in getting more information about the research group (15 mentions), as well as a positive impact on strategic publication planning (13 mentions, see Fig. 3). Fewer mentions were received by the categories performance comparison, tool for steering, and interdisciplinarity (all 6 mentions). Relatively few mentions fell on the categories quick overview (4 mentions), motivation to publish (4 mentions) and new information (3 mentions), interdisciplinary tolerance grade, bootlicking, visualization of expertise and hierarchy, and argumentation basis (1 mention each).

Retrospective analysis in this case means trying to understand how a work group has performed and cooperated over a certain period of time. Connections likes subgroups that publish together become apparent and development of work-group foci can be seen, when looking at clusters in a graph. In particular how the team has developed over time becomes visible. *Information about the team* refers to gaining insights into team make up in the current situation. It allows seeing who currently works with whom, how intensively they cooperate and who might have been left out. *Strategic publication planning* was seen as a benefit by the participants, which means that seeing your publications behavior could give you input on how to find co-authors that might benefit future publications either in regard to personal or institutional development.

Some of the participants mentioned *performance comparison*, which means the tracking of how much a person published in regard to how much publishing is expected from him. The visualization as a *tool for steering* was also mentioned by the participants, meaning that providing such a visualization to a team-leader, could

allow him to actively manage publication efforts by giving him both insights into how publishing in his group works and whether requirements are met.

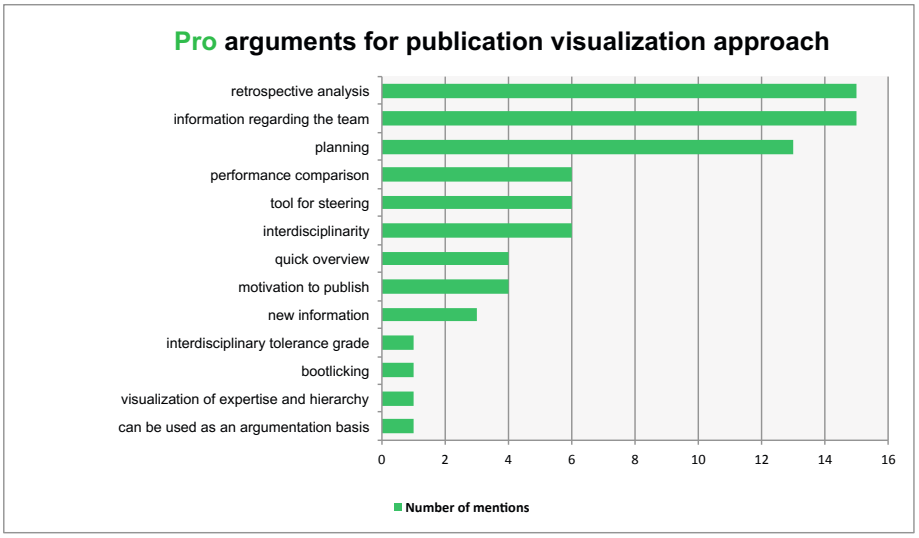

Fig. 3. Number of mentions of pro arguments for the presented visualization approach. Total number of mentions $n = 76$.

4.2 Con Arguments

The strongest concern mentioned was *missing background information* on the publications or authors (21 mentions, see Fig. 4). By simply looking at a visualization that focuses on co-authorship, personal properties such as half-time working and time at an institute are masked, as well as publication properties like an impact factor or relevance. The second strongest concern was the *negative influence on workflow* (9 mentions), which encompassed concerns like triggering competition between team-members or a general disconnect from team members. The thirst most mentioned concern was the *bad legibility* (7 mentions) of the produced graph. In particular researchers that had published multiple articles were overwhelmed by the sheer amount of lines and texts that appear in the visualization.

A general concern against the approach of visualizing publications in such a manner (i.e. not using citation data) was mentioned, that the visualization did not contain *information on quality* of the articles addressing the problem of the quality-quantity dilemma. The visualization only acknowledges quantity. A similar concern is raised by *publications are just one aspect of performance* (5 mentions), which highlights the concern, that not all useful effort researchers do is found in publications (e.g. grant applications, personal development, teaching efforts). This aspect was also mentioned by participants that saw larger graphs

(i.e. members of larger work-groups), as participants found the *scope of the visualization* to large for sensible interpretation (4 mentions).

Some additional concerns were mentioned: Among them the worry, that the visualization will have *no impact on performance* (3 mentions), that distances between authors are inappropriate, and the general question, whether steering is desirable (both 2 mentions). The least concern was found for the categories *willingness for cooperation* is unclear, *publications outside the network* are not visible, *publishing and function team* are not the same thing, and *fake authorships* (1 mention each).

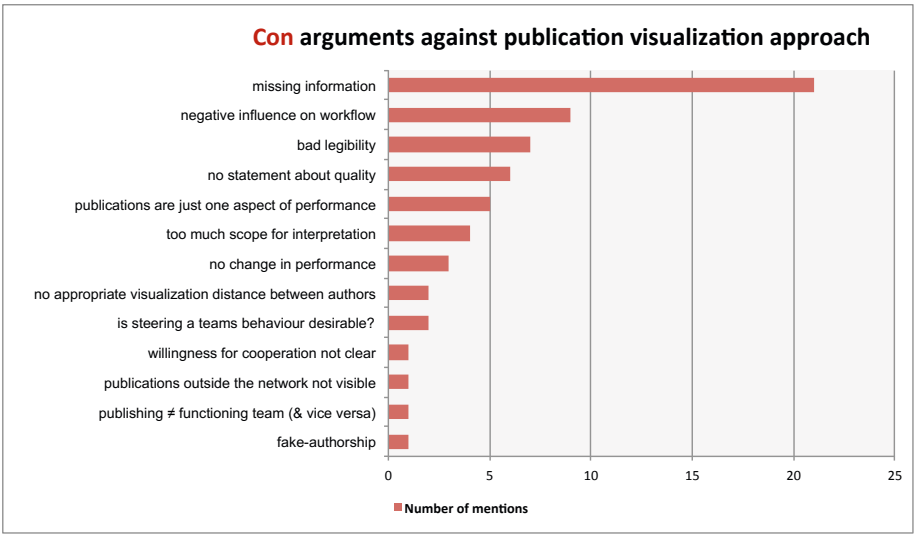

Fig. 4. Number of mentions of con arguments against the presented visualization approach. Total number of mentions $n = 63$.

4.3 Additional Requirements

Participants were also asked about wishes and added requirements for the development of the visualization tool. Among them were aspects like increasing *usability* to allow masking of nodes in order to make the text more legible. Participants also wished for a *time axis* allowing the user to move smoothly through the development of the network. Users wished that *additional information* such as impact-factors or journal names, could be attached to the nodes, to improve understanding on how and where the work group publishes. *Filter functions* such as hiding professors, other workgroups/institutes were seen as necessary, as most of the work actually happens between the individual researchers on a lower hierarchy level. One user wished to have *publishing thresholds*, that would change

the color of an author if the author had not published enough according to the threshold. *Exporting the data* into a specific citation style was wished, so that it could be used at personal/institutional web pages. Allowing to visualize other *sociometric data* was also mentioned, such as proximity (who sits in the same office). A key requirement was also that clicking onto nodes should directly take the user to a *profile page* of the author.

5 Discussion

In the interview study presented in this paper we looked at arguments for and against the usage of a publication visualization on a social platform. In general the reception of the visualization was positive, as it was seen as an enabling tool to improve the cooperation in an interdisciplinary team. These results confirm earlier findings [15] and underline the importance of integrating the user in the development process when visualizing sensitive data such as performance indicators.

It is interesting to see that even researchers with no formal training in bibliometric evaluation immediately see concerns with performance evaluation according to bibliometric data — rightly so. The development of the tool must ensure, that additional information of the bibliometric data must be integrate-able into the visualization to change the focus of the visualization from performance evaluation to cooperation understanding. On the other hand it was concerning to witness how quickly participants were able to draw "conclusions" from the graph, probably confirming existing prejudices. The presented graphs include mostly young researchers with few publications, thus making statistically reliable results almost impossible to infer. Drawing conclusions on the "performance" of a future researcher from citation data of two or three publications is particularly careless as citations do not occur normally distributed over publications. Citations show power law distributions meaning few papers get lots of citations (while most get very few) and thus typical statistics like means or variances are not meaningful or even well defined [17].

From the findings we conclude, that bibliometric social network visualization can be helpful when presented in a social portal. Nonetheless it is necessary to educate future users about the interpretability of bibliometric data and stress that performance evaluations can only be executed by trained bibliometricians. In order to still reap the benefits of our visualization approach different visibility styles will be used for different purposes. On a publicly visible level only aggregated data about larger workgroups will be visible to prevent over-interpretation of individual data. Researchers will have a private view on their individual publishing track record and their co-authors so they can still reap the benefits of understanding who they work with. Furthermore we want to allow users to enable sharing of their data selectively, so they can allow other researchers to view their "private" network with their consent. This should prevent negative impacts on workflow. We also plan to give courses on bibliometric evaluation to project leaders and people in higher hierarchy positions as a requirement to access more data to leverage the visualization for steering and planning.

6 Limitations and Future Research

The limitations of this research address four main aspects. The primary aspect is sample selection, as the sample size is always a compromise of effort and validity. The other aspects are the prototypic realization of the tool, problems with bibliometric accuracy, and the social media integration.

The results presented in this study were generated by addressing twenty-two researchers from two projects at the RWTH Aachen University. And effort was made to select participants that are good representatives of the different levels of hierarchy and different structures within the projects. This selection process plus the self-selction bias (no participants were obligated to take part) might nonetheless have preferred participants that are more open to solutions like our visualization tool.

The visualization tool was presented as a prototype, not integrated into the social portal. One reason for this approach, was not to present sensitive data to an unsuspecting audience before assessing the perceived sensitivity of the data. This might have influenced the perception of applicability of the approach, because participants could not experience the context of the tool. An open questions remains, whether users will still be as positive regarding the tool, if they see it connected to their individual profile and visible to 180 colleagues. In order to ensure acceptance, a gradual introduction of visualizations accompanied by quantitative evaluation is planned for the production version of the platform.

In order to ensure comparability within an interdisciplinary setting, we will perform a study assessing the importance of individual bibliometric indicators to the communities within the research cluster. It is necessary to regard the high level of individuality to ensure, comparability is maintained or at least addressed consciously to prevent premature judgement.

Since the visualization is integrated into a social platform all aspects that matter in social media are important immediately as well. Aspects of data privacy, establishment of business process (how to deal with under- or over-performing), etiquette (how do we talk about different publishing behaviors) are important as well but were not explicitly addressed in the interview study. Prior research [18,19] has shown that user diversity factors are highly important when looking at aspects of data disclosure and establishing an etiquette for online communication. These aspects need also be addressed explicitly in a scientific context, to ensure that the second strongest concern of *negative influence on the workflow* does not occur.

Acknowledgments. We would like to thank the anonymous reviewers for their constructive comments on an earlier version of this manuscript. Furthermore we would like to thank our student assistants Tatjana Hamann and Juliana Brell for their support in conducting the interviews. The work related to the project "Integrative Production Technology in High Wage Countries" has been funded by the Excellence Initiative of the German federal and state governments.

References

1. Nissani, M.: Ten cheers for interdisciplinarity: The case for interdisciplinary knowledge and research. The Social Science Journal 34(2), 201–216 (1997)
2. Repko, A.F.: Interdisciplinary research: Process and theory. Sage (2008)
3. Jooß, C., Welter, F., Leisten, I., Richert, A., Schaar, A.K., Calero Valdez, A., Nick, E., Prahl, U., Jansen, U., Schulz, W., et al.: Scientific cooperation engineering in the cluster of excellence integrative production technology for high-wage countries at RWTH Aachen University. In: Proceedings of the ICERI 2012, pp. 3842–3846 (2012)
4. Schuh, G., Aghassi, S., Caler Valdez, A.: Supporting technology transfer via web-based platforms. In: Proceedings of PICMET 2013 Technology Management in the IT-Driven Services (PICMET), pp. 858–866. IEEE (2013)
5. Garfield, E.: Citation indexes for science. a new dimension in documentation through association of ideas. International Journal of Epidemiology 35(5), 1123–1127 (2006)
6. Harzing, A.W., Van der Wal, R.: Google scholar: the democratization of citation analysis. Ethics in Science and Environmental Politics 8(1), 61–73 (2007)
7. Nisonger, T.E.: Citation autobiography: An investigation of isi database coverage in determining author citedness. College & Research Libraries 65(2), 152–163 (2004)
8. Hicks, D.: The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. Scientometrics 44(2), 193–215 (1999)
9. Leydesdorff, L., Shin, J.C.: How to evaluate universities in terms of their relative citation impacts: Fractional counting of citations and the normalization of differences among disciplines. Journal of the American Society for Information Science and Technology 62(6), 1146–1155 (2011)
10. Delgado López-Cózar, E., Robinson-García, N., Torres-Salinas, D.: Manipulating google scholar citations and google scholar metrics: simple, easy and tempting (2012)
11. Calero-Medina, C., Noyons, E.: Combining mapping and citation network analysis for a better understanding of the scientific development: The case of the absorptive capacity field. Journal of Informetrics 2(4), 272–279 (2008)
12. Holzinger, A., Ofner, B., Stocker, C., Calero Valdez, A., Schaar, A.K., Ziefle, M., Dehmer, M.: On graph entropy measures for knowledge discovery from publication network data. In: Cuzzocrea, A., Kittl, C., Simos, D.E., Weippl, E., Xu, L. (eds.) CD-ARES 2013. LNCS, vol. 8127, pp. 354–362. Springer, Heidelberg (2013)
13. Calero Valdez, A., Schaar, A.K., Ziefle, M., Holzinger, A., Jeschke, S., Brecher, C.: Using mixed node publication network graphs for analyzing success in interdisciplinary teams. In: Huang, R., Ghorbani, A.A., Pasi, G., Yamaguchi, T., Yen, N.Y., Jin, B. (eds.) AMT 2012. LNCS, vol. 7669, pp. 606–617. Springer, Heidelberg (2012)
14. Bastian, M., Heymann, S., Jacomy, M.: Gephi: An open source software for exploring and manipulating networks (2009)
15. Schaar, A.K., Calero Valdez, A., Ziefle, M.: Publication network visualization as an approach for interdisciplinary innovation management. In: 2013 IEEE International Professional Communication Conference (IPCC), pp. 1–8. IEEE (2013)
16. Glaser, B.G., Strauss, A.L.: The discovery of grounded theory: Strategies for qualitative research. Transaction Books (2009)

17. Redner, S.: How popular is your paper? an empirical study of the citation distribution. The European Physical Journal B-Condensed Matter and Complex Systems 4(2), 131–134 (1998)
18. Schaar, A.K., Calero Valdez, A., Ziefle, M.: The impact of user diversity on the willingness to disclose personal information in social network services. In: Holzinger, A., Ziefle, M., Hitz, M., Debevc, M. (eds.) SouthCHI 2013. LNCS, vol. 7946, pp. 174–193. Springer, Heidelberg (2013)
19. Calero Valdez, A., Kathrin Schaar, A., Ziefle, M.: Personality influences on etiquette requirements for social media in the work context. In: Holzinger, A., Ziefle, M., Hitz, M., Debevc, M. (eds.) SouthCHI 2013. LNCS, vol. 7946, pp. 427–446. Springer, Heidelberg (2013)