Who are we talking about?: identifying scientific populations online

Julie M. Birkholz[1], Shenghui Wang[2], Paul Groth[3], and Sara Magliacane[3]

[1] Network Institute, VU University Amsterdam
j.m.birkholz@vu.nl
[2] OCLC Research
shenghui.wang@oclc.org
[3] Computer Science Department, VU University Amsterdam
{p.t.groth, s.magliacane}@vu.nl

Abstract. In this paper, we begin to address the question of which scientists are online. Prior studies have shown that Web users are only a segmented reflection of the actual offline population, and thus when studying online behaviors we need to be explicit about the representativeness of the sample under study to accurately relate trends to populations. When studying social phenomena on the Web, the identification of individuals is essential to be able to generalize about specific segments of a population offline. Specifically, we present a method for assessing the online activity of a known set of actors. The method is tailored to the domain of science. We apply the method to a population of Dutch computer scientists and their co-authors. The results when combined with metadata of the set provide insights into the representativeness of the sample of interest.
The study results show that scientists of above average tenure and performance are overrepresented online; suggesting that when studying online behaviors of scientists we are commenting specifically on behaviors of above average performing scientists. Given this finding, metrics of Web behaviors of science may provide a key tool for measuring knowledge production and innovation at a faster rate than traditional delayed bibliometric studies.

1 Introduction

Traditionally, science is assessed using bibliometric techniques – indicators/metrics used to classify scientific output (e.g. publications) by performance and innovation, such as citation scores, or journal impact factors. Such methods rely on publication traditions including citations in a socially regulated environment. Communication and exchange of knowledge is also happening on the Web. The use of the Web as a virtual environment for interaction and exchange provides a ground for assessing impact through the study of traceable behaviors. Shifting research behaviors to the the Web in multiple domains exposes more and more diverse processes of knowledge production and communication. Behavior online is traceable. Consequently scientists behaviors on the Web provide an additional, arguably complimentary, set of information traces to study science.

With the general rise of Internet use, an increasing portion of researchers' work takes place online via e-mail exchange, accessing online bibliographic databases, blogging, collaborating through e-science tools, as well as general Web usage. A number of studies have begun to explore the online behaviors of science communities (see [16]). These research projects suggest that at face value similar communication conventions, such as citation of academic articles occur on both blogs [5] and Twitter [14]. The rise in these online activities suggests that an increasing portion of knowledge production and discussion is occurring on the Web, in parallel to traditional practices of knowledge dissemination through academic publication, conferences and the like. Theses online platforms have consequently been seen as a new terrain for exploring knowledge production as well as science assessment, through assessing the "total impact" of a scientists work article [17].

The validity of these metrics remains debatable. This lack of established validity is an issue for the generation and testing of theories of research behavior and scientific development based on this data. One challenge in the validation of what these metrics represent is to determine who we are actually talking about. We pose the question: in science, who is represented online? And in what manner?; working to answer the level of validity that Web metrics providing in commenting on populations in science. We describe a method based on a combination of Social Science theory and Computer Science methods to evaluate the representativeness of a set of actors on the Web, compared to an offline population. The method focuses on understanding whether a known sample is active on a variety of social platforms (e.g. Twitter, Mendeley, LinkedIn). Using a focused crawling approach, we examine a population of Dutch computer scientists and their co-authors presence on the Web. We show that:

- a relatively low percentage of these scientists are verifiably active online;
- those who are active on social websites are likely to be active on multiple sites;
- and that the scientists who are online are largely high performing.

The rest of the paper is organized as follows. We begin with a survey of studies of science online, highlighting the known differences between on- and offline communication. We then describe the method itself, which is followed by a description of the results of our study on Dutch computer scientists. Finally, we discuss the results and conclude.

2 Science Online

The practice of science is a practice of communication; an act of dissemination of knowledge to a set of peers. Scientific knowledge is disseminated through scientific publication in journals, conference proceedings and books. Consequently, these outputs are used as measures of knowledge production in science, through bibliometrics. Bibliographic records, through the use of repositories and databases, thus provide a wealth of knowledge on the system of science. These studies not only shed light on performance or innovation through classification of outputs, but also the collaborative (co-authorship) behaviors of scientists (see work related to [3]; and network studies from [12]) undertaken to produce knowledge.

The Web is a platform that expands on these practices of knowledge dissemination. As Wellman [22] suggested online behaviors mimic actual social behavior. There are a number of outlets for scientists that cater to scientists [16]. An increasing amount of studies have specifically explored scientists online from how online behaviors spark creativity in science [4], differences between online and offline behaviors of scientists [11], field differences in using computer mediated communication [21], as well as the function of virtual communities in science [2]. Consequently we know scientists are online, and have variety of options for disseminating knowledge and interacting with peers outside of the traditional/formal publication of knowledge to hard copy text.

These research studies have set a path of inquiry about the effects of knowledge dissemination via the Web. A short list of tools exist that aid in further conceptualizing and understanding these online social behaviors in science which include methods to track online readership [20], and impact metrics [13]. It is the use of tools that has been explored as way to complement bibliometric assessment.

Where publication trends are applicable for a field or discipline the use of the Web is not a uniform nor required practice within science. Studies of the Web show that subsegments of populations are more likely to be on the Web, specifically younger ones [7]; correspondingly certainly not everyone is on the Web, and particularly not everyone in science but the composition is unknown. It is these behaviors that are of interest for this study, as they are the individual actions of the scientists that have emerged outside of traditional communication system.

The rise of scientists' use of the Web as a platform for sharing knowledge presents a number of methodological questions of validity and reliability to accurately reflect on how Web behaviors, in contrast to publication records, which include all active scientists within fields of practice, these individual behaviors present a case where activities are voluntary, relate to scientists/knowledge production. By validity we refer to – how to generalize a sample to a larger population, how to generalize across settings, or both generalize across and within samples. Two aspects of validity need to be considered in science – external and internal. Internal validity relates to the suitability of measures for the population being examined. External validity refers to if results can infer causal mechanisms for an entire population [10]; giving an indication of the generalizability of the research to a specific population. The external validity is dependent on the population that a study aims to generalize about.

In order to accurately comment on scientists' practices online one must be explicit about the validity of the sample. Specifically the representativeness of studies of these growing online behaviors of scientists on the Web remains a question to not only the description of the Web but also the implications that we can infer from the sample on the Web. This validity is critical for generalizing and connecting research findings to other knowledge products. In this study, we work to define a method and test the results; developing a marker which defines the reliability of a Web sample in relation to a specific greater population of scientists under study.

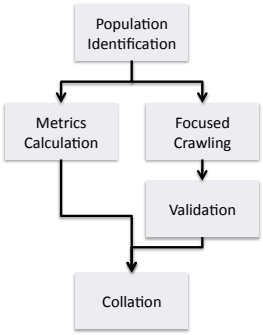

Fig. 1. Method Overview

3 Methodology

In this study a method is developed to identify individuals on a number of Web platforms. This description is followed by an explanation of the application to a known population. Statistical analysis is done on the results to reflect on the representativeness of the specific sample under study. We begin this section with a general description of the Web crawler method.

3.1 Method description

Identifying scientists online could be achieved in one of two ways: query a web platform for users or starting from a known sample, that you know you want to generalize about and identify online. Both choices have a number of draw backs. The general query does not ensure we can correctly identify names of scientist to place them into a population (American scientist, historian, and so forth). The second option is privy to that scientists identify themselves online in a logical/disambiguable way where we can link scientists through their full names John Smith is JohnSmith on Twitter and not thewhistler. Thus, both choices of sample select suffer from disambiguation issues. We do not aim to discuss disambiguation techniques in this paper, as that is a field in itself, but rather acknowledge that there are a number of techniques, of which techniques we integrate here.

Since we are interested in identifying scientists on multiple platforms and aim to compare them in some way we choose for the second sample technique.

A Web crawler is suited to identify individual online. Although the crawler described could be used for other communities it is particularly suited for investigating scientists on the Web. The key steps of the method and the data flow between those steps are shown in Figure 1. We describe each of these steps in turn.

Population identification The method starts with identifying an already known population of scientists. Here, we are looking for a precise disambiguated set of individuals. An example list consists of names as well as other metadata such as affiliation. There are a number of mechanisms for gathering such data including using institutional directories, retrieving author lists from bibliographic databases (e.g. the Web of Science)

or through membership lists from academic societies. Such lists provide a reasonable picture of offline membership of the population under question.

Collecting standard science metrics Once a population list is obtained, the sample needs to be characterized in more detail. One could imagine a number of ways to characterize the population of scientists including age, gender, institutional type (e.g. teaching verses research university), tenure, and so forth. Here, we classify scientists according to standard metrics used in science studies, namely, the h-index and citation scores due to the difficultly of automatically collecting traditional variables of individuals from the Web. Regardless, the measures of performance provide information from which we can infer about their likely age/tenure through citations as well as performance and or value of knowledge to a community through citation score and h-index. These two standard measures of scientists knowledge impact provide a representative manner to reflect on the population in terms of activity within the particular scientific community. These statistics also provide a basis for comparison when looking at the population online. Science metrics can be obtained from a number of databases such as Web of Science, Google Scholar or Scopus.

Focused crawling In this study, we investigate the online behaviors of scientists' own enterprise/ individual actions. This implies that we are not searching in online bibliographic databases for evidence of publications, or their academic institutional pages; but rather that we are isolating the existence of online activity on the social Web including: blogs, micro-blogging, and activity on social platforms.

In science, blogs are often used as an alternative dissemination space for knowledge, whether presenting new knowledge, ideas or research, or sharing information. Micro-blogging tools, such as Twitter,[1] are used by academics for sharing academic links [15]. LinkedIn[2] is a known professional social-networking site used by academics as well as other professionals. Mendely[3] is a bibliographic bookmarking service that aids scientists in organizing academic publications and links, as well as sharing them through profile libraries. Slideshare[4] is a site used to upload presentations, providing an outlet for scientists to disseminate lectures and presentations. The diverse services offered among these Web platforms provide an outlet to incorporate/consider the multiple forms of knowledge dissemination on the Web.

To obtain information about the online behavior of individuals on these various sites, we developed a focused Web crawler. This crawler takes as input the list of persons from the population identification stage. It automatically performs the following process.

For each scientist, the Web crawler first goes over her/his homepage and searches for evidence of online presence such as links to her/his blog, "follow me" links for LinkedIn, or Twitter, as well as entries in Mendeley and Slideshare. If these activities are not mentioned on the personal homepage, the crawler individually searches LinkedIn, Twitter, Mendeley and Slideshare to check whether she/he exists on these sites.

[1] http://twitter.com/
[2] http://www.linkedin.com/
[3] http://www.mendeley.com/
[4] http://www.slideshare.net/

The crawler takes the scientist's name as the search string and submits the query to each of these websites, specifically searching for people. If the search returns zero hits, then we consider that the scientist does not have an account on the websites. Very often, the search returns multiple hits. This is due to the way the search strings are handled by the different websites. For example LinkedIn and Twitter only return accounts whose full names match exactly the target scientist, while Mendeley and Slideshare return the accounts whose name contains either the first name or the last name of the target scientist. For the latter situation, we filter out the accounts whose full names are not the same as the target scientist.

Apart from the binary information about whether one scientist is present in the above mentioned social platforms, data can also be collected on the activity of the scientist. In this case, we focus on Twitter as an example. On Twitter we can also gather information about following and friends (followers) counts in Twitter to say something additional about Web behaviors of identified scientists.

The result of this step of the method is a list of possible accounts on these online sites corresponding to a given person in the input population.

Validation The results of the focused crawl provide some evidence that a particular individual is present online but because multiple hits may occur we need to perform further validation to determine whether indeed an individual is present.

To validate the data a more detailed comparison is carried out based on the metadata of the returned results and the descriptive information of the target scientist. The query results usually contain some metadata of the returned accounts, such as the ID, full name, occupation, location, etc. We further check whether the location information of the query results match the location information of the institution where the target scientist works given in the population list. This is accomplished through the use of the Yahoo PlaceFinder web service.[5] This service provides the latitude and longitude as well as the country and city information for both institutions and the returned accounts. If the account is in the same country as the target scientist, we consider this account belongs to the target scientist. If multiple accounts still are in question after the location check (scientists who share the same country) we consider that the target scientist exists in this Web platform without further distinguishing which accounts belongs to her/him. This is an approach in favor of recall.

Collation The results of validation are collated together with the information obtained from the metrics calculation step.For each scientists in the population we have information about their membership in a community, their performance, tenure, as well as their participation online. Based on this information, standard statistical measures can then be run to analyze how the online activity relates to the any number of factors within standard science metrics.

[5] http://developer.yahoo.com/geo/placefinder/

3.2 Method Application

We apply the above method to a population of presently active computer scientists working in nine Dutch academic research institutions and their co-authors.

To perform the population identification step, we obtained a list of Dutch computer scientists from the Dutch NARCIS - National Academic Research and Collaborations Information System database.[6]

We expanded this list by querying the Digital Bibliography and Library Project (DBLP) [9] – an online bibliographic database for the field of Computer Science with publication streams from a number of top rated journals, conference proceedings and books within the field. We queried for all source scientists and co-authors from January 2007 (the year after Twitter's inception) to March 2011. This query returned in total 4984 individual scientists. They represent a list of all active scientists connected to Dutch computer scientists via co-authorship. This source list is the population we are interested in this study/that we can generalize about.

After the population identification step, traditional science metrics (h-index and citation score) for each scientist were collected. This data was acquired through ArnetMiner [19] a search and mining services of Computer Science researchers which includes semantic data on names, contact information, homepage, and additional traditional scienteometric statistics. Arnetminer identified 4590 scientists (394 less than the 1st query); providing the metadata to aid in depicting how this sample represents computer scientists online. The use ArnetMiner rather than the Web of Science or Scopus is important for this specific population because of the better coverage of Computer Science related publications in broader databases [1].

The population data was input into the focused Web crawler and collated with the data from ArnetMiner. The specific results for this population in terms of each website are given in the next section.

4 Results

In this research, we conducted bivariate Pearson's correlations to explore the external validity of a population of computer scientists. Correlations allow us to determine relationships between two or more variables. Results from such tests identify both significant relationships and the direction of relationships (positive or negative); thus providing the researcher with evidence to suggest, but not confirm, a causal relationship. Note, that in the results table a ** denotes statistical significance at the 0.01 level, and * denotes statistical significance at the 0.05 level.

Our findings of the sample of Dutch computer scientists and their co-authors from January 2007 to March 2011 show a relatively low percentage of scientists online. As displayed in Table 1, each of the web platforms are analyzed describing the frequency and percentages represented in the sample under study, with the value 1 representing a positive identification by the web crawler of individuals, and 0 representing a negative identification by the crawler. For LinkedIn 81.5% of the sample can be identified, with 18.5% not found on the platform. In Mendeley, we see 89.8% of scientists not identified

[6] http://www.narcis.nl/

		LinkedIn		Mendeley		Slideshare	
		Frequency	Percent	Frequency	Percent	Frequency	Percent
Valid	0	851	18.5	4121	89.8	4522	98.5
	1	3739	81.5	469	10.2	68	1.5
	Total	4590	100.0	4590	100.0	4590	100.0

		blog		Twitter	
		Frequency	Percent	Frequency	Percent
Valid	0	4438	96.7	4503	98.1
	1	152	3.3	87	1.9
	Total	4590	100.0	4590	100.0

Table 1. Descriptive statistics of how many author names can be found on the various services and whether those names can be validated.

on this site, and 10.2% confirmed. On Slideshare we find 98.5% not identified and 1.5% identified. Twitter identification is 1.9% and 98.1% not identified. Within this sample only 3.3% of scientists are identified as having a blog, and 96.7% not identified. Thus, for the sample of Dutch computer scientists and their co-authors only a small share is identified on all Web platforms, with the largest shares on LinkedIn and Mendeley.

Results show that computer scientists who are active on the Web are likely to be active on multiple sites (see Table 2. Correlations of platforms). Within this population the existence on LinkedIn is related to being identified on Mendeley (r=0.124), Slideshare (r=0.054), Twitter (r=0.058), and having a blog (r=0.057). The strongest relationship is existence on both LinkedIn and Mendeley. This strong positive correlation holds the same for the relationships of Mendeley to the Web platforms of Slideshare (r=0.072), Twitter (r=0.064) and blogs (r=0.090), with the strongest relationship to Mendeley being the existence of a blog. Slideshare identification also strongly correlates with the existence on other platforms: Twitter (r=0.128) and blog (r=0.088); suggesting the strongest relationship between the use of Slideshare and Twitter. The use of Twitter and a blog also has a strong positive relationship (r= 0.394). Overall, those active on Web platforms have tendencies to be active on multiple sites.

To further contextualize the use of online platforms, we present additional data from Twitter (see Table 3 Correlations to Twitter activity). The correlations results show that those on Twitter have both high numbers of followers and following (r=0.777). We also investigate the relationships to the followers and following and the identification on other Web platforms: there is no significant relationship between LinkedIn (following: r=0.025; followers: r=0.026), although significant positive relationships exist between Mendeley (following: r=0.091; followers: r=0.108), Slideshare (following: r=0.143; followers: r=0.121) and blog identification (following: r=0.186; followers: r=0.176). The strongest relationship between following and followers is with Slideshare and blog use. This suggests a positive relationship between the number of followers and following and activity on other sites.

In order to further understand what is driving the positive relationships observed, it is necessary to also investigate the relationships of traditional performance measures to describe what extend this population online is generalizable (externally valid) to (Dutch) Computer Science as a field. These results are presented in Table 4. Results

		LinkedIn	Mendeley	Slideshare	blog	Twitter
LinkedIn	Pearson Correlation	1	.124**	.054**	.057**	.058**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	4590	4590	4590	4590	4590
Mendeley	Pearson Correlation	.124**	1	.072**	.090**	.064**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	4590	4590	4590	4590	4590
Slideshare	Pearson Correlation	.054**	.072**	1	.088**	.128**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	4590	4590	4590	4590	4590
blog	Pearson Correlation	.057**	.090**	.088**	1	.394**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	4590	4590	4590	4590	4590
Twitter	Pearson Correlation	.058**	.064**	.128**	.394**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	4590	4590	4590	4590	4590

Table 2. Correlations of platforms

		Twitter	Twitter following	Twitter friends
h-index	Pearson Correlation	.118**	.083**	.058**
	Sig. (2-tailed)	.000	.000	.000
	N	4590	4590	4590
total-citation	Pearson Correlation	.077**	.065**	.032*
	Sig. (2-tailed)	.000	.000	.031
	N	4590	4590	4590
LinkedIn	Pearson Correlation	.058**	.025	.026
	Sig. (2-tailed)	.000	.092	.073
	N	4590	4590	4590
Mendeley	Pearson Correlation	.064**	.091**	.108**
	Sig. (2-tailed)	.000	.000	.000
	N	4590	4590	4590
Slideshare	Pearson Correlation	.128**	.143**	.121**
	Sig. (2-tailed)	.000	.000	.000
	N	4590	4590	4590
blog	Pearson Correlation	.394**	.186**	.176**
	Sig. (2-tailed)	.000	.000	.000
	N	4590	4590	4590
Twitter	Pearson Correlation	1	.375**	.399**
	Sig. (2-tailed)		.000	.000
	N	4590	4590	4590
Twitter Lfollowing	Pearson Correlation	.375**	1	.777**
	Sig. (2-tailed)	.000		.000
	N	4590	4590	4590

Table 3. Correlations to Twitter activity

		LinkedIn	Mendeley	Slideshare	blog	Twitter
h-index	Pearson Correlation	.070**	.021	.038*	.154**	.118**
	Sig. (2-tailed)	.000	.159	.011	.000	.000
	N	4590	4590	4590	4590	4590
total-citations	Pearson Correlation	.034*	-.009	.004	.103**	.077**
	Sig. (2-tailed)	.020	.528	.805	.000	.000
	N	4590	4590	4590	4590	4590

Table 4. Performance measures

show that those online are largely top ranking scientists; with the higher h-index the more likely to be found on LinkedIn (r=0.070), Slideshare (r=0.038), Twitter (r=0.118) and having a blog (r=0.154). There is no signification relationship between identification on Mendeley and a high h-index score (r=0.021). A number of these relationships are also confirmed in regards to citation score (number of citations), which is used as a measure for performance and a proxy for tenure. A positive and significant relationship exists between citation score and LinkedIn (r=0.034), Twitter (r=0.077), and identification of a blog (r=0.103). Results suggest that among this community of computer scientists the measuring of Web behaviors of scientists' own enterprise is representative of dynamics of scientists who have both a higher tenure and higher performance.

5 Discussion

Before discussing the results, it is important to reflect on the limitations of the Web crawler. The implementation of the Web crawling tool, which takes the names as input and automatically searches the presence of these people on the Web, greatly increases the amount of people who can be analyzed; thus, providing a more reliable extension to manual tests of validity of specific communities. Limitations of the Web crawler include: disambiguation issues, and API constraints and limits. The most common disambiguation issue is the lack of meaningful IDs that match full names. Search APIs also present some limits to searches for scientists. APIs sometimes use OR instead of AND to increase the recall, which presents a problem in quickly and reliably locating a name among the results of crawler, thus, requiring additional knowledge about individuals. LinkedIn returns entities whose names contain the full string of searched names; while Mendeley and Slideshare return people whose names contain either first name OR last name. In the development of the Web crawler, this was overcome through the integration of geolocation data to identify individuals within the returned set. Additionally, some APIs have limits to queries per hour, which constrains the speed of the crawler. The constraints of the Web crawler potentially affect the low identification of the sample online. Further techniques could be developed in the Web crawler to provide more certainty about scientists presence on these sites. In particular, we are looking at building profiles of scientists based on publications and matching these to profiles produced from websites. Such an approach may help in increasing both the recall and precision of the method [6].

This test showed that the largest percentages of scientists can be identified on LinkedIn and Mendeley, with much lower identification on Slideshare, Twitter and blogs. Although this could be related to the techniques of the Web crawler, we suggest that it is

rather associated to the services provided on these sites. LinkedIn provides a networking tool for professionals to connect on the Web, which we argue reflects traditional communication patterns of scientists [8] staying in touch whether through email, phone contact or face-to-face interaction. Mendeley provides an online bibliographic bookmarking tool that again scientists would be in need of whether on or offline to categorize and organize publications. The other three platforms: Slideshare, Twitter, and blogs are forms of modern communication and thus new ways of disseminating knowledge. They are not innate to the knowledge dissemination practices of the past several hundred years unlike interacting with others (LinkedIn), and reading, reflecting and reacting to new knowledge in the field (Mendeley).

The significant positive relationships observed in this activity sample on multiple sites give us reason to hypothesize that scientists are using Web platforms in their work, thus providing further support to our previous speculation of tendencies in using specific platforms that facilitate traditional knowledge production. To confirm this, further research should be completed to explore the online presence of other fields and samples of scientists, shedding light on the overall prevalence of online activity in science. Additionally, the results from Twitter bring to light a possible feedback effect of using multiple Web platforms. This effect should be explored further to address the mechanism of using such Web platforms, but also how visibly increasing on one platform relates to other Web behaviors. Consequently longitudinal research of scientists' online activities would provide insight into the effect of the use of multiple Web platforms.

This study has shown that when using Web data we oversample dynamics of top scientists; bringing to light the importance of considering validity questions of Web data to study social phenomena. Thus, when talking about implications of altmetrics [18] or analyzing behavior on these social media sites we need to be explicit about who we can generalize about and how these reflect to greater patterns in science. If this oversampling of top scientist holds true for other fields, the use of Web data may provide a reliable, faster tool in measuring, predicting and understanding trends in science; compared to delayed bibliometric analysis of top scientists.

6 Conclusion

Our results present a depiction of life on the Web for the field of Computer Science. From an analytics perspective, we have worked to develop a method that provides a tool for reflecting on population as to reliably provide a level of external validity (generalizability) to a greater community of actors. Additionally, it emphasizes the importance and continued need of interdisciplinary research to assess such questions.

In summary, we propose that when measuring the scientific impact and contribution of ones work on the Web it is necessary to be clear about the level of external validity of these Web activities in order to better infer trends in science. The method described here is a first step to achieving this goal.

References

1. J. Bar-Ilan. Web of science with the conference proceedings citation indexes: the case of computer science. *Scientometrics*, 83(3):809–824, 2010.

2. L. Colazzo, A. Molinari, and N. Villa. From e-learning to "co-learning": the role of virtual communities. In M Kendall and B Samways, editors, *IFIP International Federation for Information Processing 281*, pages 329–338, 2008.
3. D.J. de Solla Price. Networks of scientific papers. *Science*, 149:510–515, 1965.
4. K. Dunbar. How scientists think: Online creativity and conceptual change in science. In T. B. Ward, S. M. Smith, and S. Vaid, editors, *Conceptual structures and processes: Emergence, discovery and change*, American Psychological Association, pages 461–493, Washington, DC, 1997.
5. P. Groth and T. Gurney. Studying scientific discourse on the web using bibliometrics: A chemistry blogging case study. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, Raleigh, NC: US, April 26-27th 2010.
6. Thomas Gurney, Edwin Horlings, and Peter van den Besselaar. Author disambiguation using multi-aspect similarity indicators. *Scientometrics*, pages 1–15. DOI: 10.1007/s11192-011-0589-1.
7. K. Hampton, L. Sessions-Goulet, L. Rainie, and K. Purcell. Social networking sites and our lives. *Pew Research Center*, June 16 2011.
8. B. Latour and S. Woolgar. *Laboratory Life: the Social Construction of Scientific Facts*. Sage Publications, Los Angeles, 1979.
9. Michael Ley. Dblp - some lessons learned. *PVLDB*, 2(2):1493–1500, 2009.
10. J. W. Lucas. Theory-testing, generalization, and the problem of external validity. *Sociological Theory*, 21:236253, 2003.
11. Peter Mika. Social networks and the semantic web. In *Web Intelligence*, pages 285–291, 2004.
12. M.E.J. Newman. The structure of scientific collaboration networks. In *Proceedings of the National Academy of Sciences 98*, pages 404–409, 2001.
13. C. Neylon and S. Wu. Article-level metrics and the evolution of scientific impact. *PLoS Biol*, 7(11: e1000242), 2009.
14. J. Priem and K Costello. How and why scholars cite on twitter. In *Proceedings of the 73rd ASIS&T Annual Meeting*, Pittsburgh, PA, USA, 2010.
15. J. Priem, K. Costello, and T. Dzuba. Prevalence and use of twitter among scholars. In *Metrics 2011: Symposium on Informetric and Scientometric Research. Poster*, New Orleans, LA, USA, October 2011.
16. J. Priem and B.M. Hemminger. Scientometrics 2.0: Toward new metrics of scholarly impact on the social web. *First Monday*, (7), 2010.
17. J Priem, C Parra, H Piwowar, P Groth, and A Waagmeester. Uncovering impacts: a case study in using altmetrics tools. *Workshop on the Semantic Publishing SePublica 2012 at the 9th Extended Semantic Web Conference*, pages 1–5, 2012.
18. J. Priem, D. Taraborelli, P. Groth, and C. Neylon. Alt-metrics: A manifesto, (v.1.0). http://altmetrics.org/manifesto, 26 October 2010.
19. Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 990–998, New York, NY, USA, 2008. ACM.
20. D. Taraborelli. Readermeter: Crowdsourcing research impact. *Academic Productivity*, 2010. Retrieved April 5, 2011, from: http://www.academicproductivity.com/2010/readermeter-crowdsourcing-research-impact/.
21. J. P. Walsh and T. Bayma. Computer networks and scientific work. *Social Studies Science*, 26:661703, 1996.
22. B. Wellman and M. Gulia. Virtual communities as communities. In M.A. Smith and P. Kollock, editors, *Communities in Cyberspace.*, Routledge, New York, 1999.