Posted, Visited, Exported: Altmetrics in the Social Tagging System BibSonomy

Daniel Zoller[a,*], Stephan Doerfel[b], Robert Jäschke[c], Gerd Stumme[b,c], Andreas Hotho[a,c]

[a] *Data Mining and Information Retrieval (DMIR) Group, University of Würzburg, Am Hubland, 97074 Würzburg, Germany*
[b] *Interdisciplinary Research Center for Information System Design (ITeG) & Knowledge and Data Engineering (KDE) Group, University of Kassel, Wilhelmshöher Allee 73, 34121 Kassel, Germany*
[c] *L3S Research Center, Appelstraße 4, 30167 Hannover, Germany*

Abstract

In social tagging systems, like Mendeley, CiteULike, and BibSonomy, users can post, tag, visit, or export scholarly publications. In this paper, we compare citations with metrics derived from users' activities (altmetrics) in the popular social bookmarking system BibSonomy. Our analysis, using a corpus of more than 250,000 publications published before 2010, 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. Machine learning classifiers predict whether the number of citations that a publication receives in a year exceeds the median number of citations in that year, based on the usage counts of the preceding year. In that setup, a Random Forest predictor outperforms the baseline on average by seven percentage points.

Keywords: altmetrics, scholarly impact, social bookmarking, collaborative tagging

1. Introduction

Scholarly impact is traditionally measured in scores computed from counting citations to publications. However, citation counts have the drawback of being only available long after an article has been published – simply because it takes time to write and publish new articles with a corresponding reference. Citations can be used as impact indicators, but they do not help researchers to find papers which will be important for their discipline in the near future, for instance, 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. Similar 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 forth – 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".[1] The Altmetrics Manifesto (Priem et al., 2011) 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

*Corresponding author.
Email addresses: zoller@informatik.uni-wuerzburg.de (Daniel Zoller), doerfel@cs.uni-kassel.de (Stephan Doerfel), jaeschke@L3S.de (Robert Jäschke), stumme@cs.uni-kassel.de (Gerd Stumme), hotho@informatik.uni-wuerzburg.de (Andreas Hotho)

[1] http://altmetrics.org/about/

missing evidence as one of the (current) disadvantages of altmetrics. Following this demand, in this paper, we focus on the social web system BibSonomy,[2] a bookmarking tool for publication references and investigate correlations between usage metrics and citations, as well as their predictive potential 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[3]) 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. All data are available within the system such that neither external sources nor the full texts of the publications are required for the computation of the metrics. Particularly, we go beyond previous work in this area

i) by explicitly comparing behavioral features to citations that occur in the near future (within one year) instead of comparing to all citations,
ii) by using algorithms from machine learning to estimate the predictive potential over future citations,
iii) 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
iv) by adding the use case of BibSonomy to the set of social web systems that have been investigated as possible sources for altmetrics.[4]

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 in topics without using external information about them, unlike previous works, where only publications of a particular conference or journal were chosen (Bar-Ilan et al., 2012; Haustein & Siebenlist, 2011; Li et al., 2012; Priem et al., 2012; Saeed et al., 2008) or where external classification systems were used to partition articles into disciplines (Haustein et al., 2014a; Thelwall & 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 to understand the form of impact that is measured by usage metrics. Thus, our research questions are the following:

i) *Can we detect a usage bias towards highly cited publications (in terms of correlations) in the large BibSonomy corpus that spans various disciplines and contains publications of different quality?*

ii) *Can the bookmarking system's tagging feature be used to create topic-homogeneous subsets in which altmetrics exhibit higher correlations with citations than the full corpus?*

iii) *Can BibSonomy data, in particular the observable traces of user behavior, be used to predict future citations (citations that occur* after *the observed usage of a publication)*?

In the remainder of the paper, we first discuss related work in Section 2 and then explain the extracted features and discuss expectations and limitations in Section 3. We describe the dataset in Section 4 and present the analysis and findings regarding the above research questions in Section 5. Section 6 concludes the paper.

2. 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.

[2] `http://www.bibsonomy.org/`

[3] `http://academic.research.microsoft.com/`

[4] We are aware of only one other altmetric investigation that included BibSonomy data: Haustein & Siebenlist (2011) computed metrics on the journal level, whereas we focus on the article level.

2.1. Scientometrics and Altmetrics on the Web

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 only being available 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 other facts, 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., 2011) as well as the altmetrics workshops.[5] As 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 $r = 0.397$). 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. However, for each system the share of articles covered by their corpus 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 1.4 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 instance, 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. Only a small or no correlation was observed between citations and social measures[6] derived from Academia.edu, Mendeley, or ResearchGate.

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.

[5] `http://altmetrics.org/workshop2011/`, `http://altmetrics.org/altmetrics12/`, `http://altmetrics.org/altmetrics14/`

[6] Social measures can be computed as altmetrics on the author level; Ortega (2015) consider followers and followees per author.

2.2. 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 & Siebenlist, 2011), on the level of authors (Bar-Ilan et al., 2012), and on the level of publications (Bar-Ilan et al., 2012; Li et al., 2012; Priem et al., 2012; Saeed et al., 2008).

Haustein & 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. For their analysis, 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 from 0.304 to 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 (Bar-Ilan et al., 2012; Li et al., 2012; Saeed et al., 2008) 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, for example, from a particular conference or journal. In contrast to these restrictions, our corpus includes any publication that users have posted to BibSonomy.

Thelwall & 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 (2015); Thelwall & Wilson (2014) and Brzezinski (2015) confirmed that publication corpora of various disciplines indeed possess different statistical properties. The dataset in BibSonomy is heterogeneous as there is no mechanism in BibSonomy to explicitly assign a publication to a 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 & Thelwall, 2014; Mohammadi et al., 2015; Thelwall & 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 & 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 & 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 paper, 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. 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 paper, however, we investigate particularly the correlation between usage metrics and citations that occur in the future, that is, in the year after some evidence of activity in BibSonomy.

Temporal aspects of altmetrics have been considered by Thelwall & 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 explanation, they suggest, is that in the early years Mendeley readerships were 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, that is, 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 next subsection) and must be kept in mind when interpreting the results.

2.3. Scholarly Bookmarking in BibSonomy

BibSonomy is an open, publicly available social publication management and sharing system on the web where users store, tag (annotate), share, and manage publications and website bookmarks. Details on the system's architecture can be found in (Benz et al., 2010). In this paper, we ignore bookmarks to websites and focus only on publications. In BibSonomy, they can be used to compile publication lists (e.g., for current research, for seminars, or as "to-read" list) and they can be exported into various literature citation formats. A publication post (or publication bookmark) usually contains the user who owns it (who entered it), some tags, and some meta data of the publication, including title, authors, and year of publication, as well as some other optional information like the publication venue, publisher address, and so forth.

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[7] and almost 100 million documents.[8] Probably the most similar to BibSonomy is CiteULike, which has more than eight million articles[9] at the time of writing. The current publicly available dataset[10] contains 145,744 users. BibSonomy currently has about

[7] `http://blog.mendeley.com/elsevier/mendeley-and-elsevier-2-years-on/` (accessed August 28, 2015)
[8] `https://www.mendeley.com/compare-mendeley/` (accessed August 28, 2015)
[9] `http://www.citeulike.org/` (accessed August 28, 2015)
[10] `http://www.citeulike.org/faq/data.adp` (accessed August 28, 2015)

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.

BibSonomy has been subject to various scientific investigations using publicly available data dumps. For instance, Bogers (2009) compared and optimized different recommendation algorithms to suggest relevant publications to users, while Jäschke et al. (2008) investigated the recommendation of tags. Sun et al. (2013) used BibSonomy data to validate an agent-based model for the development of scientific disciplines using the collaboration behavior of researchers. Recently in (Doerfel et al., 2014b), for the first time, the weblogs of a bookmarking system have been analyzed to study user behavior, particularly, how social the social tagging system is. Research on the usage of the posted publications (Doerfel et al., 2014a) revealed small correlations between a publication's popularity in posts and popularity in requests. It is therefore reasonable to investigate these two popularities as possible impact metrics. In fact, they are very similar to the metrics *post* and *req*, that we will discuss in this work (see Section 3.1).

3. Alternative Metrics in BibSonomy

Before we present the datasets in the next section, here, we address the setup of our experiments and discuss expectations and limitations of our study.

3.1. Behavioral Metrics and Setup

In our experiments, we use six different metrics as indicators for a publication's impact:

i) The metric $post(p)$ counts how often a publication p was bookmarked in BibSonomy. This is the same metric that was used in previous literature on BibSonomy or other systems (Bar-Ilan et al., 2012; Haustein & Siebenlist, 2011; Li et al., 2012; Priem et al., 2012; Saeed et al., 2008).

ii) With $view(p)$, we denote how often a publication p has been viewed in BibSonomy (e.g., the publication's details page or a page with all posts about this publication from different users).

iii) We denote with $exp(p)$ the number of times a publication p has been exported into citation formats (e.g., BibTeX or EndNote).

iv) Since BibTeX is the most often requested export format on BibSonomy we additionally use the metric $exp_{Bib}(p)$ to count exports of p to that format.

v) 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.

vi) 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 behavior 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 behavior 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 a given year, we denote the number of citations to that 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 ($cit^{+0}, cit^{+1} \leq cit$).

3.2. Methodology

In this work, we conduct two kinds of analyses: We measure the correlation between behavioral metrics and citations, and we investigate the predictive potential of these metrics over citations in the future, that is, in the year after the observed activity in BibSonomy. For correlations, we report Pearson's correlation coefficient r, as well as Spearman's ranking correlation ρ. Because the latter has the advantage that it is suitable for non-linear relationships, we focus on ρ for the upcoming discussions. 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, that is, entities with their features and their classes. The trained model can then be evaluated by applying it to an unlabeled dataset.

In our classification setting, the entities are publication-year pairs, and the features are the observed behavioral metrics in that year. Given a set of publications together with their usage metrics[11] per year, we use all publication-year pairs (p, y), where for publication p at least one of the used metrics was positive in year y, that is, 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 τ. That is, 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 τ 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[12] 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:

$$(a, 2007): \quad post = p_{2007},\ view = v_{2007}, \quad cit^{+1} = c_{2008}$$
$$(a, 2008): \quad view = v_{2008},\ exp = e_{2008}, \quad cit^{+1} = c_{2009}\,.$$

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:

$$(a, 2009): \quad exp = e_{2009}, \quad cit^{+1} = c_{2010}\,.$$

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 median 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.

As classifiers we selected implementations of *Random Forest* (Breiman, 2001) and *SVMs* with different kernels (Cortes & Vapnik, 1995), covering the two best classifier families at the moment (see, for instance, (Fernández-Delgado et al., 2014)). For Random Forest we used the implementation of the R-package randomForest[13] with its standard configuration and with 100 repetitions per experiment. The SVMs were chosen from the SVM^{light} package[14] using a radial and a polynomial kernel, again with default parameters.

[11] In Section 5.3, we will use the metrics *post*, *exp*, and *view*, following the results in Section 5.2.

[12] 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.

[13] `http://cran.r-project.org/web/packages/randomForest/index.html`

[14] `http://svmlight.joachims.org`

3.3. Expectations and Limitations

Before we report the results of our analysis in the next section, we discuss our expectations and also limitations of this study.

3.3.1. Expectations

In Section 3.1 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_{Bib} 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 Sections 1 and 2, we have explained that the dataset in this study is less restricted than those of previous studies, that is, it contains arbitrary publications contributed by BibSonomy's users instead of only publications from a particularly popular venue. Furthermore, BibSonomy is relatively small compared to systems used in previous studies (see Section 2.3). 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, etc.). It is well known in bibliometrics that both the scientific discipline and the venue are influential factors for a publication's probability of receiving citations (Bornmann & Daniel (2008) survey a variety of studies regarding these influences). We therefore expect much lower correlations than those reported in the previous experiments mentioned in Section 2 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; and it is unknown how that will influence the observable correlations. Our assumption is that users of BibSonomy manage the publications they plan to cite with BibSonomy. However, in general, the reasons why users choose to post a publication are diverse – from saving and organizing, over highlighting and annotating to (self-)marketing (Haustein et al. (2015)). In BibSonomy we noticed that many users store meta data of their own work, for example, 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.

3.3.2. 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). Orduña-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 & Thelwall (2014)). Similarly, to (Waltman & 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 behavioral 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 Section 2.3) 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 & Costas, 2012), that apply for BibSonomy and for our study as well: As data are user-generated, they are error-prone and depend 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 5.2, where we use tags to produce topic-focused subsets of publications. 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.

4. Dataset

For our experiments, we combine meta data 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.

4.1. BibSonomy

The dataset used in this paper 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.[15] 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. We make anonymized datasets of logs and posts available to researchers.[16]

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 are obtained by crawling the web and extracting information from publications. According to Khabsa & 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 & Giles, 2014; Orduña-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.

[15] Both datasets also include data from later years, but for our experiments, we had to restrict them (see Section 4.3).

[16] `http://www.kde.cs.uni-kassel.de/bibsonomy/dumps/`

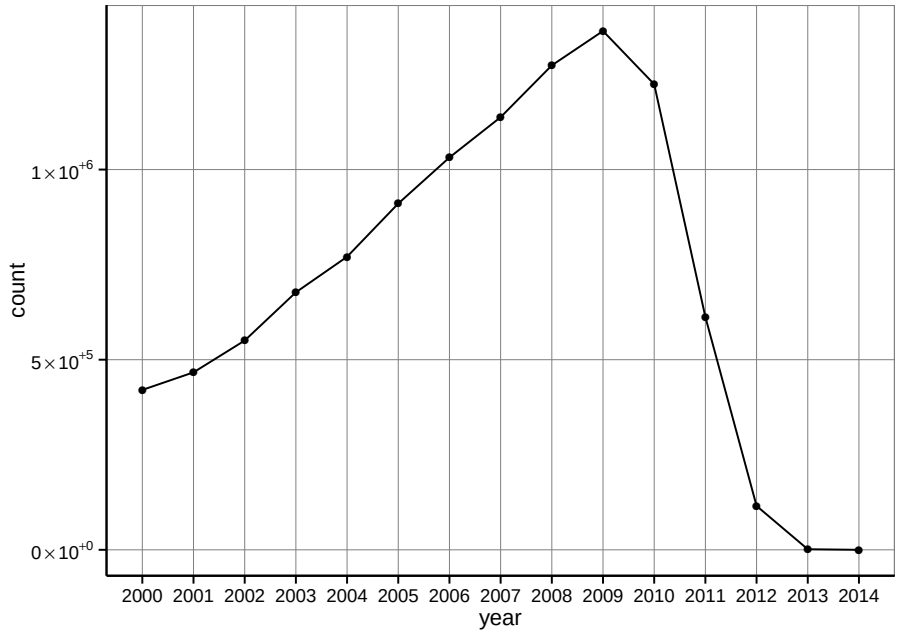


Figure 1: Citations to publications in BibSonomy according to MAS, distributed over the years. For better readability only the last 15 years are displayed.

4.3. Matching between BibSonomy and MAS

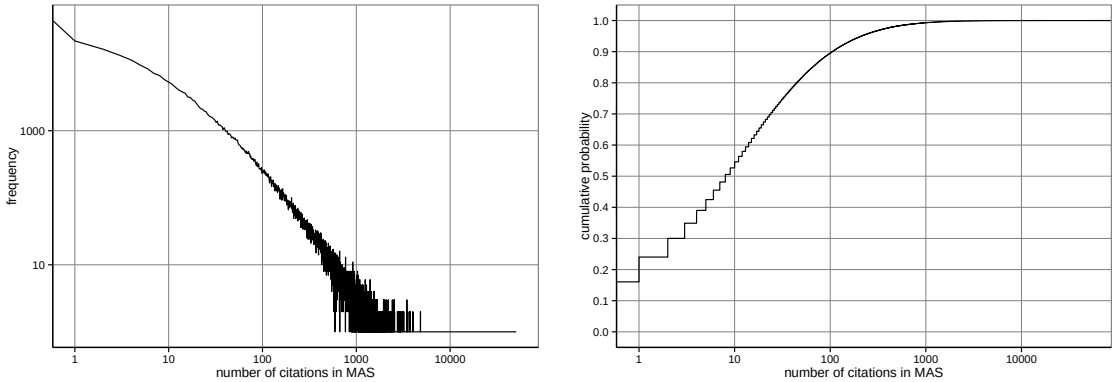
Dealing with publication meta data is often tedious: The data from BibSonomy is user-generated and thus prone to contain spelling errors and incomplete attributes, making it difficult to match publications. However, also the data about references in publications – as it is collected and extracted by 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 & 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 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, 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.[17] Publications that appeared after 2010 are rarely indexed (see below).

[17] `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 2: The frequency distribution and the cumulative distribution function of the number of citations recorded in MAS for publications in BibSonomy.

- 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 forth.
- Not all publications in BibSonomy actually have been published – this includes preprints, manuscripts, or bachelor/master theses.
- 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.

Orduña-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 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 Orduña-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 Orduña-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.

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 2a. Similar frequency distributions were observed in previous works, for instance, by Redner (1998), on other citation datasets. 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 2b. 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), that was also used in previous studies of citation distributions (Albarrán & Ruiz-Castillo, 2011; Brzezinski, 2015).

Table 1: Correlation between different behavioral features in BibSonomy and the number of citations of a publication. The upper right triangle shows Pearson's r, the lower left triangle shows Spearman's ρ. For each correlation value, the upper and lower bound of the 99% confidence interval is reported. Correlations are computed over all publications in the dataset.

	post	*view*	*exp*	exp_{Bib}	*req*	*tag*	*cit*
post	1	.644 [.636, .651]	.638 [.629, .647]	.633 [.623, .642]	.446 [.437, .455]	.330 [.323, .337]	.181 [.174, .187]
view	.322 [.311, .332]	1	.725 [.716, .733]	.705 [.696, .714]	.656 [.649, .662]	.322 [.310, .334]	.091 [.079, .104]
exp	.317 [.303, .332]	.429 [.414, .443]	1	.988 [.988, .988]	.742 [.735, .748]	.279 [.263, .294]	.157 [.141, .173]
exp_{Bib}	.328 [.314, .343]	.417 [.401, .431]	.955 [.953, .956]	1	.722 [.714, .729]	.277 [.261, .293]	.160 [.143, .176]
req	.325 [.315, .336]	.912 [.910, .914]	.663 [.654, .671]	.634 [.625, .643]	1	.213 [.201, .225]	.072 [.060, .084]
tag	.277 [.270, .285]	.272 [.259, .284]	.237 [.220, .253]	.242 [.225, .258]	.267 [.255, .278]	1	.036 [.028, .044]
cit	.199 [.193, .206]	.098 [.086, .111]	.122 [.106, .138]	.120 [.104, .137]	.098 [.086, .110]	.014 [.007, .022]	1

The optimal fit has parameters $\alpha = 2.59$ and $x_{\min} = 808$. Since 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 also consider the second (local) optimum: $\alpha = 2.34$ for $x_{\min} = 303$. The exponents α of both fits are lower than reported by both Brzezinski (2015) and Albarrán & 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 log-normal 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 & Wilson (2015), who observed for 45 medical sub-fields that log-normal distributions are better fits to citation counts than power laws. Thelwall & Wilson (2015) further showed that for these disciplines a hooked power law is an even better fit than the log-normal distribution.

Moreover, the high x_{min} values suggest the presence of a "hooked power law" (Thelwall & 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 & 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 α is lower than those measured by Thelwall & Wilson (2014) for citation distributions of various research areas (the minimum α there was $\alpha = 2.51$).

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 α is lower than previously observed both in the power-law and the hooked power-law fit).

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 5.2 and attempt the prediction of citations in Section 5.3.

5.1. Correlations on the Full Corpus

Our first experiment is similar to those in the literature mentioned in Section 2, 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 1 shows Pearson's r and Spearman's ρ 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. The

Table 2: For each behavioral metric, the correlations (Pearson's r and Spearman's ρ) between the metric and citations in the same year (cit^{+0}) and citations in the following year only (cit^{+1}) with their corresponding upper and lower bounds of the 99% confidence interval. 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.

	Pearson's r				Spearman's ρ				
metric	cit^{+0}		cit^{+1}		cit^{+0}		cit^{+1}		N
post	.201	[.196, .207]	.199	[.194, .205]	.163	[.157, .168]	.158	[.153, .164]	194,012
view	.117	[.107, .127]	.117	[.107, .127]	.139	[.129, .149]	.145	[.135, .155]	64,355
exp	.171	[.158, .185]	.166	[.152, .179]	.150	[.136, .164]	.148	[.135, .162]	33,989
exp_{Bib}	.172	[.158, .186]	.167	[.153, .181]	.151	[.137, .165]	.151	[.137, .165]	31,985
req	.063	[.054, .072]	.062	[.053, .071]	.141	[.132, .150]	.145	[.136, .154]	77,018
tag	.038	[.034, .042]	.034	[.030, .038]	.067	[.063, .071]	.059	[.055, .063]	399,502

confidence intervals of all correlation coefficients are very small (less than ± 0.02). As expected (see Section 3.3) 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 BibTeX exports (exp_{Bib}) makes little difference – 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_{Bib}, 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, i.e., where the publication was used at least once in that year, according to the metric. Table 2 shows the results: All metrics except *tag* (as before) exhibit medium correlations with citations in the near future ($0.14 \leq \rho \leq 0.16$). As in the previous experiment, the confidence intervals at 99% are very narrow. 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 & 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.

5.2. Correlations for Popular Topics

It is well known in scientometrics (see Section 3.3) 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 & 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" vs. "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

Table 3: For each behavioral metric, the correlations (Pearson's r and Spearman's ρ) 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.

	Pearson's r		Spearman's ρ		
metric	cit^{+0} ±sd.	cit^{+1} ±sd.	cit^{+0} ±sd.	cit^{+1} ±sd.	N ±sd.
post	0.324 ±0.094	0.331 ±0.097	0.290 ±0.065	0.310 ±0.075	1,551.0 ±950.5
view	0.157 ±0.080	0.169 ±0.078	0.198 ±0.063	0.242 ±0.071	1,119.6 ±516.8
exp	0.280 ±0.140	0.275 ±0.134	0.257 ±0.072	0.270 ±0.074	664.7 ±316.5
exp_{Bib}	0.280 ±0.149	0.276 ±0.142	0.263 ±0.066	0.276 ±0.067	634.7 ±303.7
req	0.192 ±0.098	0.200 ±0.094	0.216 ±0.064	0.248 ±0.071	1,260.0 ±579.6
tag	0.121 ±0.091	0.116 ±0.089	0.143 ±0.088	0.128 ±0.099	4,890.7 ±3,651.5

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 A (Table A.5 for cit^{+0} and Table A.6 for cit^{+1}). Table 3 is similar to Table 2, only instead of correlations on the full corpus they are now measured on each of the 30 small ones individually and then averaged. We note that these (unweighted) average correlations are all much higher than those observed on 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 CiteULike (see Section 2), 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 4 shows, for each tag, Pearson's r and Spearman's ρ (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 1, 2, and 3), exp_{Bib} is a sub-metric of *exp* with almost perfect correlation (Table 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 4). For twelve tags, we observe average correlations larger than or equal to $\rho = 0.300$ (rounded).

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.

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 3.2.: 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 3.2): Random Forest and two SVMs, one with a 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 50% which would be achieved by random guessing. Therefore, we also compare to the latter as *Baseline Random*.

Figure 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 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: 8.68×10^{-7} for Major and 5.77×10^{-8} for Random).

Table 4: For each of BibSonomy's 30 most popular tag stems, average correlations between the behavioral metrics and citation counts in the future (within the next year), together with the standard deviation (sd.), ordered by their Spearman's correlation ρ. 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. ±sd.	ρ avg. ±sd.
structur (224)	0.305 ±0.100	0.431 ±0.031
folksonomi (278)	0.218 ±0.091	0.381 ±0.025
web20 (234)	0.319 ±0.082	0.379 ±0.054
tag (294)	0.314 ±0.082	0.365 ±0.045
collabor (269)	0.311 ±0.118	0.356 ±0.054
inform (397)	0.337 ±0.150	0.333 ±0.052
web (409)	0.332 ±0.122	0.320 ±0.043
cluster (241)	0.272 ±0.061	0.316 ±0.068
network (384)	0.271 ±0.113	0.300 ±0.040
social (354)	0.150 ±0.098	0.297 ±0.060
commun (339)	0.292 ±0.100	0.295 ±0.004
algorithm (241)	0.251 ±0.055	0.295 ±0.049
data (269)	0.208 ±0.021	0.289 ±0.047
ontolog (357)	0.256 ±0.083	0.284 ±0.024
search (224)	0.255 ±0.080	0.256 ±0.047
system (382)	0.218 ±0.121	0.255 ±0.042
semant (380)	0.215 ±0.066	0.253 ±0.040
learn (309)	0.168 ±0.077	0.252 ±0.067
analysi (358)	0.211 ±0.086	0.249 ±0.026
theori (324)	0.349 ±0.151	0.242 ±0.028
model (460)	0.172 ±0.060	0.235 ±0.037
knowledg (261)	0.180 ±0.025	0.233 ±0.020
languag (219)	0.174 ±0.047	0.220 ±0.046
evalu (285)	0.247 ±0.071	0.212 ±0.020
internet (226)	0.179 ±0.057	0.209 ±0.053
softwar (277)	0.284 ±0.098	0.208 ±0.013
comput (317)	0.506 ±0.080	0.204 ±0.070
design (289)	0.484 ±0.088	0.202 ±0.063
process (254)	0.172 ±0.059	0.175 ±0.036
manag (298)	0.105 ±0.053	0.168 ±0.039
full corpus	0.151 ±0.012	0.151 ±0.007

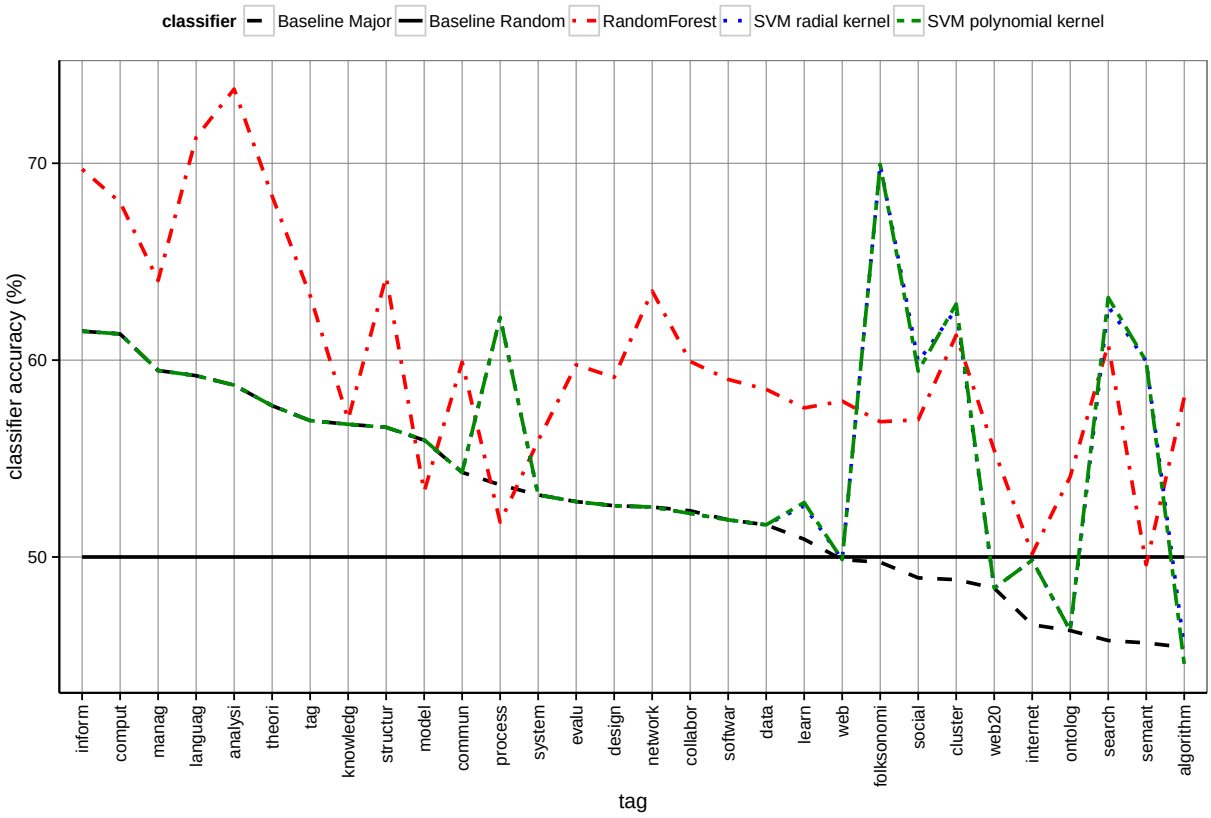


Figure 3: Classification accuracy for the 30 tag-induced publication subsets for three classifiers and two baselines. Note that the diagrams for the two SVMs are almost indiscernible as they often yield identical results. The subsets are ordered by the share of entities in the test set that belong to the class that is most frequent in the training set.

On average Baseline Random is exceeded by 9.97 percentage points and Baseline Major by 7.13 percentage points. 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: 1.86×10^{-8} for Major and 5.59×10^{-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 5.00×10^{-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 paper.

6. Conclusion

With the analyses in Section 5, we can answer the three research questions from the introduction:

i) 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 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.

ii) We saw that using tags to group publications into topics successfully increased the correlation 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.

iii) We found small, yet significant predictive power in an experiment where usage features (bookmarks, views, and exports) of one year were used by machine learning classifiers to predict whether the number of citations that a publication receives in the next year exceeds the median number of citations in that next year. 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 3.3.2 and the restriction to works published before 2010 that was induced by the available data in MAS, our results must be interpreted with care. Also, 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 recommendation, 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. The observed correlations and predictive power even if small suggests that the usage intensity is related to (future) citation impact but also confirms that these altmetrics are not the same as citation impact. They can be used as indicators for impact that will only later be acknowledged (formally) by being cited, and therefore BibSonomy is another potential source for altmetrics. Since none of the measures are 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).

Future Work. The results in this paper 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, for instance, for webmasters who want to present predictions of future relevance in their system.

Another aspect worth analyzing is that of diversity within these metrics, for example, 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 analyze which impact the age of publications has on the accuracy of predictions, (iii) to investigate deeper, which features contribute best to successful predictions, and (iv) to adapt and optimize classifiers to yield better predictions.

References

Albarrán, P., & Ruiz-Castillo, J. (2011). References made and citations received by scientific articles. *Journal of the American Society for Information Science and Technology*, *62*, 40–49. doi:10.1002/asi.21448.

Bar-Ilan, J., Haustein, S., Peters, I., Priem, J., Shema, H., & Terliesner, J. (2012). Beyond citations: Scholars' visibility on the social web. In *Proceedings of 17th International Conference on Science and Technology Indicators* (pp. 98–109).

Benz, D., Hotho, A., Jäschke, R., Krause, B., Mitzlaff, F., Schmitz, C., & Stumme, G. (2010). The social bookmark and publication management system BibSonomy. *The VLDB Journal*, *19*, 849–875. doi:10.1007/s00778-010-0208-4.

Bogers, T. (2009). *Recommender systems for social bookmarking*. Ph.D. thesis Tilburg University The Netherlands.

Bornmann, L. (2014a). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, *8*, 895–903. doi:10.1016/j.joi.2014.09.005.

Bornmann, L. (2014b). Validity of altmetrics data for measuring societal impact: A study using data from altmetric and F1000Prime. *Journal of Informetrics*, *8*, 935–950. doi:10.1016/j.joi.2014.09.007.

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*, 45–80. doi:10.1108/00220410810844150.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. doi:10.1023/A:1010933404324.

Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, *57*, 1060–1072. doi:10.1002/asi.20373.

Brzezinski, M. (2015). Power laws in citation distributions: Evidence from Scopus. *Scientometrics*, *103*, 213–228. doi:10.1007/s11192-014-1524-z.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*, 661–703. doi:10.1137/070710111.

Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning*, *20*, 273–297.

Costas, R., Zahedi, Z., & Wouters, P. (2014). 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*, *66*, 2003–2019. doi:10.1002/asi.23309.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*, 171–176. doi:10.1145/363958.363994.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Doerfel, S., Zoller, D., Singer, P., Niebler, T., Hotho, A., & Strohmaier, M. (2014a). Evaluating assumptions about social tagging – A study of user behavior in BibSonomy. In *Proceedings of the 16th LWA Workshops: KDML, IR and FGWM, Aachen, Germany*. CEUR-WS.org.

Doerfel, S., Zoller, D., Singer, P., Niebler, T., Hotho, A., & Strohmaier, M. (2014b). How social is social tagging? In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion* WWW Companion '14 (pp. 251–252). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. doi:10.1145/2567948.2577301.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*, 3133–3181.

Fu, T., Song, Q., & Chiu, D. (2014). The academic social network. *Scientometrics*, *101*, 203–239. doi:10.1007/s11192-014-1356-x.

Golder, S. A., & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, *32*, 198–208. doi:10.1177/0165551506062337.

Haley, M. R. (2014). 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*, 1079–1084. doi:10.1002/asi.23080.

Haustein, S., Bowman, T. D., & Costas, R. (2015). Interpreting "altmetrics": viewing acts on social media through the lens of citation and social theories. arXiv:1502.05701.

Haustein, S., Larivière, V., Thelwall, M., Amyot, D., & Peters, I. (2014a). Tweets vs. Mendeley readers: How do these two social media metrics differ? *it - Information Technology*, *56*, 207–215. doi:10.1515/itit-2014-1048.

Haustein, S., Peters, I., Sugimoto, C. R., Thelwall, M., & Larivière, V. (2014b). Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *Journal of the Association for Information Science and Technology*, *65*, 656–669. doi:10.1002/asi.23101.

Haustein, S., & Siebenlist, T. (2011). Applying social bookmarking data to evaluate journal usage. *Journal of Informetrics*, *5*, 446–457. doi:10.1016/j.joi.2011.04.002.

Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., & Stumme, G. (2008). Tag recommendations in social bookmarking systems. *AI Communications*, *21*, 231–247. doi:10.3233/AIC-2008-0438.

Kaur, J., JafariAsbagh, M., Radicchi, F., & Menczer, F. (2014). Scholarometer: A system for crowdsourcing scholarly impact metrics. In *Proceedings of the 2014 ACM Conference on Web Science* (pp. 285–286). New York, NY, USA: ACM. doi:10.1145/2615569.2615669.

Khabsa, M., & Giles, C. L. (2014). The number of scholarly documents on the public web. *PLoS ONE*, *9*, 1–6.

Li, X., Thelwall, M., & Giustini, D. (2012). Validating online reference managers for scholarly impact measurement. *Scientometrics*, *91*, 461–471. doi:10.1007/s11192-011-0580-x.

Mas-Bleda, A., Thelwall, M., Kousha, K., & Aguillo, I. (2014). Do highly cited researchers successfully use the social web? *Scientometrics*, *101*, 337–356. doi:10.1007/s11192-014-1345-0.

Mohammadi, E., & Thelwall, M. (2014). Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. *Journal of the Association for Information Science and Technology*, *65*, 1627–1638. doi:10.1002/asi.23071.

Mohammadi, E., Thelwall, M., Haustein, S., & Larivière, V. (2015). Who reads research articles? An altmetrics analysis of Mendeley user categories. *Journal of the Association for Information Science and Technology*, *66*, 1832–1846. doi:10.1002/asi.23286.

Orduña-Malea, E., Ayllon, J. M., & andd Emilio Delgado López-Cózar, A. M. (2014). Empirical evidences in citation-based search engines: Is Microsoft Academic Search dead? arXiv:1404.7045.

Ortega, J. L. (2015). Relationship between altmetric and bibliometric indicators across academic social sites: The case of CSIC's members. *Journal of Informetrics*, *9*, 39–49. doi:10.1016/j.joi.2014.11.004.

Peters, I., Haustein, S., & Terliesner, J. (2011). Crowdsourcing in article evaluation. In *Proceedings of the 2011 ACM Conference on Web Science* (pp. 1–4).

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2015). Research data explored: Citations versus altmetrics. In *Proceedings of the 15th International Society of Scientometrics and Informetrics Conference* (pp. 172–183). Bogaziçi University Printhouse.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, *14*, 130–137. doi:10.1108/eb046814.

Priem, J., Piwowar, H. A., & Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. arXiv:1203.4745.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2011). Altmetrics: a manifesto. URL: http://altmetrics.org/manifesto/.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *European Physical Journal B*, *4*, 131–134.

Saeed, A., Afzal, M., Latif, A., & Tochtermann, K. (2008). Citation rank prediction based on bookmark counts: Exploratory case study of WWW06 papers. In *Proceedings of the IEEE International Multitopic Conference* (pp. 392–397). doi:10.1109/INMIC.2008.4777769.

Sun, X., Kaur, J., Milojevic, S., Flammini, A., & Menczer, F. (2013). Social dynamics of science. *Scientific Reports*, *3*. doi:10.1038/srep0106.

Thelwall, M. (2015). Why do papers have many Mendeley readers but few Scopus-indexed citations and vice versa? *Journal of Librarianship and Information Science*, *n/a*. doi:10.1177/0961000615594867.

Thelwall, M., & Fairclough, R. (2015). The influence of time and discipline on the magnitude of correlations between citation counts and quality scores. *Journal of Informetrics*, *9*, 529–541.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PLoS ONE*, *8*. doi:10.1371/journal.pone.0064841.

Thelwall, M., & Sud, P. (2015). Mendeley readership counts: An investigation of temporal and disciplinary differences. *Journal of the Association for Information Science and Technology*, *n/a*. doi:10.1002/asi.23559.

Thelwall, M., & Wilson, P. (2014). Distributions for cited articles from individual subjects and years. *Journal of Informetrics*, *8*, 824–839. doi:10.1016/j.joi.2014.08.001.

Thelwall, M., & Wilson, P. (2015). Mendeley readership altmetrics for medical articles: An analysis of 45 fields. *Journal of the Association for Information Science and Technology*, *n/a*. doi:10.1002/asi.23501.

Voß, J., Hotho, A., & Jäschke, R. (2009). Mapping bibliographic records with bibliographic hash keys. In *Proceedings of 11. Internationales Symposium für Informationswissenschaft* (pp. 535–536). Hochschulverband Informationswissenschaft Verlag Werner Hülsbusch.

Waltman, L., & Costas, R. (2014). 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*, 433–445. doi:10.1002/asi.23040.

Wouters, P., & Costas, R. (2012). Users, narcissism and control – tracking the impact of scholarly publications in the 21st century. SURFfoundation.

Zahedi, Z., Costas, R., & Wouters, P. (2015). Do Mendeley readership counts help to filter highly cited WoS publications better than average citation impact of journals (JCS)? In *Proceedings of the 15th International Society of Scientometrics and Informetrics Conference* (pp. 16–25). Bogaziçi University Printhouse.

Appendix A. Correlations for Popular Topics

Table A.5: For each tag stem, the correlations (Pearson's r and Spearman's ρ) between the behavioral metric and citations in the same year cit^{+0} with their corresponding upper and lower bounds of the 99% confidence interval. 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).

	post		*view*		*exp*	
tag stem	Pearson's r	Spearman's ρ	Pearson's r	Spearman's ρ	Pearson's r	Spearman's ρ
model	.244 [.206, .281]	.267 [.230, .304]	.100 [.044, .155]	.164 [.109, .218]	.172 [.091, .250]	.247 [.169, .322]
web	.482 [.433, .529]	.325 [.268, .380]	.177 [.118, .235]	.223 [.165, .279]	.310 [.242, .375]	.284 [.215, .350]
inform	.365 [.308, .420]	.385 [.329, .439]	.138 [.061, .214]	.252 [.177, .324]	.517 [.438, .588]	.308 [.212, .398]
network	.420 [.374, .464]	.325 [.275, .373]	.148 [.084, .211]	.244 [.181, .304]	.228 [.148, .306]	.267 [.188, .343]
system	.397 [.346, .446]	.260 [.204, .315]	.105 [.022, .187]	.174 [.092, .253]	.158 [.046, .267]	.232 [.122, .336]
semant	.266 [.217, .313]	.276 [.228, .322]	.093 [.039, .146]	.184 [.132, .236]	.175 [.109, .239]	.216 [.152, .279]
analysi	.276 [.228, .321]	.272 [.225, .318]	.104 [.042, .166]	.215 [.155, .274]	.216 [.138, .291]	.204 [.126, .280]
ontolog	.338 [.285, .388]	.306 [.252, .357]	.128 [.072, .183]	.187 [.133, .241]	.265 [.200, .327]	.288 [.224, .349]
social	.280 [.215, .343]	.305 [.241, .367]	.058 [-.016, .133]	.183 [.110, .254]	.095 [.003, .186]	.259 [.170, .343]
commun	.254 [.183, .323]	.300 [.230, .367]	.185 [.109, .260]	.256 [.181, .328]	.439 [.350, .520]	.286 [.187, .380]
theori	.261 [.198, .322]	.204 [.139, .267]	.234 [.150, .314]	.243 [.160, .323]	.564 [.476, .642]	.263 [.147, .372]
comput	.572 [.520, .620]	.276 [.206, .343]	.404 [.316, .484]	.058 [-.042, .157]	.521 [.423, .607]	.215 [.092, .332]
learn	.289 [.234, .342]	.329 [.276, .381]	.085 [.024, .145]	.152 [.092, .211]	.144 [.060, .226]	.261 [.181, .338]
manag	.183 [.090, .273]	.212 [.119, .300]	.036 [-.074, .145]	.100 [-.010, .207]	.120 [-.020, .256]	.120 [-.020, .256]
tag	.383 [.309, .453]	.343 [.267, .416]	.182 [.102, .261]	.225 [.146, .302]	.348 [.262, .430]	.378 [.293, .457]
design	.445 [.389, .498]	.287 [.223, .348]	.382 [.315, .445]	.138 [.062, .211]	.635 [.567, .695]	.184 [.080, .285]
evalu	.195 [.105, .282]	.203 [.113, .289]	.153 [.059, .245]	.119 [.024, .211]	.342 [.227, .447]	.166 [.042, .284]
folksonomi	.343 [.253, .427]	.350 [.260, .433]	.104 [.009, .197]	.272 [.182, .358]	.212 [.108, .312]	.390 [.296, .477]
softwar	.308 [.242, .371]	.156 [.086, .224]	.150 [.078, .221]	.183 [.112, .253]	.421 [.338, .498]	.186 [.091, .278]
collabor	.461 [.395, .522]	.372 [.301, .439]	.160 [.082, .237]	.230 [.154, .304]	.278 [.184, .367]	.321 [.229, .407]
data	.250 [.188, .309]	.218 [.156, .278]	.174 [.087, .259]	.289 [.205, .369]	.225 [.109, .336]	.254 [.139, .363]
knowledg	.220 [.137, .301]	.298 [.218, .375]	.147 [.061, .231]	.226 [.141, .306]	.204 [.090, .313]	.219 [.106, .327]
process	.230 [.170, .288]	.208 [.147, .267]	.072 [-.026, .168]	.152 [.055, .246]	.114 [-.029, .253]	.137 [-.005, .275]
cluster	.339 [.274, .401]	.405 [.343, .463]	.192 [.118, .264]	.250 [.178, .320]	.282 [.186, .373]	.245 [.147, .337]
algorithm	.321 [.287, .354]	.257 [.222, .292]	.250 [.188, .310]	.256 [.194, .317]	.173 [.077, .265]	.335 [.246, .418]
web20	.391 [.299, .476]	.405 [.314, .489]	.184 [.082, .282]	.228 [.128, .324]	.368 [.260, .468]	.340 [.230, .442]
internet	.252 [.111, .384]	.209 [.066, .344]	.115 [-.047, .271]	.105 [-.056, .262]	.126 [-.077, .319]	.233 [.034, .415]
search	.319 [.240, .394]	.274 [.193, .351]	.126 [.036, .213]	.124 [.035, .212]	.251 [.145, .352]	.200 [.092, .303]
structur	.443 [.370, .510]	.402 [.327, .472]	.213 [.088, .331]	.353 [.236, .459]	.290 [.129, .435]	.451 [.308, .575]
languag	.199 [.114, .281]	.273 [.191, .352]	.100 [.004, .195]	.153 [.058, .246]	.209 [.089, .322]	.229 [.111, .341]

Table A.5: *Continued*: For each tag stem, the correlations (Pearson's r and Spearman's ρ) between the behavioral metric and citations in the same year cit^{+0} with their corresponding upper and lower bounds of the 99% confidence interval. 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).

	exp_{Bib}		*req*		*tag*	
tag stem	Pearson's r	Spearman's ρ	Pearson's r	Spearman's ρ	Pearson's r	Spearman's ρ
model	.162 [.078, .243]	.248 [.167, .326]	.119 [.067, .171]	.210 [.160, .260]	.057 [.036, .078]	.159 [.139, .179]
web	.314 [.245, .380]	.287 [.217, .355]	.217 [.162, .271]	.243 [.188, .296]	.161 [.120, .201]	.101 [.060, .142]
inform	.538 [.459, .608]	.301 [.203, .394]	.227 [.156, .295]	.278 [.209, .344]	.135 [.098, .171]	.117 [.080, .153]
network	.225 [.143, .305]	.268 [.187, .345]	.180 [.121, .238]	.241 [.183, .297]	.149 [.120, .177]	.136 [.107, .164]
system	.151 [.035, .262]	.227 [.114, .335]	.121 [.044, .197]	.179 [.103, .253]	.088 [.056, .119]	.154 [.124, .185]
semant	.161 [.094, .227]	.214 [.148, .278]	.116 [.065, .166]	.204 [.154, .253]	.152 [.118, .185]	.163 [.129, .196]
analysi	.228 [.148, .305]	.208 [.128, .286]	.140 [.082, .197]	.234 [.178, .289]	.046 [.019, .073]	.137 [.110, .163]
ontolog	.270 [.205, .334]	.277 [.211, .340]	.166 [.114, .217]	.213 [.162, .263]	.126 [.088, .164]	.132 [.094, .169]
social	.096 [.001, .189]	.261 [.170, .347]	.074 [.002, .144]	.208 [.139, .276]	.065 [.020, .109]	.110 [.065, .153]
commun	.488 [.402, .566]	.268 [.166, .365]	.258 [.188, .326]	.282 [.212, .349]	.063 [.018, .109]	.185 [.141, .229]
theori	.562 [.470, .642]	.261 [.141, .373]	.350 [.276, .419]	.261 [.184, .336]	.042 [.007, .077]	.184 [.151, .218]
comput	.497 [.393, .588]	.209 [.082, .330]	.457 [.383, .526]	.093 [.003, .182]	.058 [.021, .095]	.091 [.054, .127]
learn	.148 [.061, .233]	.290 [.208, .368]	.119 [.062, .175]	.186 [.130, .240]	.000 [-.032, .033]	.017 [-.016, .050]
manag	.083 [-.062, .225]	.155 [.011, .293]	.052 [-.051, .153]	.112 [.010, .212]	.038 [-.017, .093]	.124 [.070, .178]
tag	.367 [.280, .448]	.390 [.305, .469]	.216 [.138, .291]	.246 [.169, .320]	.333 [.276, .388]	.279 [.220, .336]
design	.650 [.581, .710]	.206 [.097, .310]	.473 [.415, .526]	.151 [.080, .220]	.048 [.012, .083]	.128 [.092, .162]
evalu	.341 [.223, .449]	.179 [.053, .300]	.204 [.116, .289]	.161 [.072, .248]	.048 [-.003, .098]	.080 [.030, .130]
folksonomi	.209 [.103, .310]	.394 [.298, .482]	.129 [.036, .219]	.297 [.210, .380]	.228 [.151, .302]	.355 [.283, .423]
softwar	.442 [.358, .518]	.199 [.103, .293]	.208 [.141, .272]	.172 [.104, .238]	.149 [.107, .190]	.020 [-.022, .062]
collabor	.265 [.169, .357]	.302 [.208, .391]	.197 [.123, .269]	.273 [.201, .342]	.146 [.095, .196]	.142 [.091, .192]
data	.224 [.104, .337]	.251 [.133, .363]	.253 [.174, .328]	.296 [.219, .369]	.125 [.092, .157]	.070 [.038, .103]
knowledg	.167 [.049, .280]	.225 [.109, .335]	.159 [.078, .237]	.219 [.140, .296]	.044 [-.008, .095]	.025 [-.026, .076]
process	.092 [-.055, .235]	.177 [.032, .315]	.098 [.006, .187]	.131 [.040, .220]	.084 [.051, .116]	.283 [.253, .312]
cluster	.253 [.153, .347]	.239 [.138, .334]	.231 [.161, .298]	.264 [.195, .330]	.162 [.121, .201]	.291 [.253, .329]
algorithm	.171 [.073, .265]	.342 [.252, .427]	.197 [.138, .254]	.275 [.218, .330]	.033 [.014, .053]	.019 [-.001, .039]
web20	.420 [.315, .515]	.367 [.257, .467]	.222 [.123, .317]	.256 [.158, .349]	.380 [.304, .451]	.314 [.234, .389]
internet	.148 [-.060, .342]	.273 [.072, .453]	.132 [-.016, .275]	.131 [-.018, .274]	.156 [.079, .230]	.035 [-.042, .112]
search	.259 [.151, .360]	.218 [.108, .322]	.061 [-.024, .146]	.143 [.058, .225]	.233 [.180, .284]	.152 [.098, .205]
structur	.267 [.103, .418]	.428 [.279, .557]	.252 [.140, .359]	.371 [.265, .467]	.260 [.218, .301]	.189 [.145, .231]
languag	.189 [.067, .306]	.233 [.113, .347]	.133 [.044, .220]	.154 [.066, .240]	.024 [-.022, .070]	.095 [.049, .141]

Table A.6: For each tag stem, the correlations (Pearson's r and Spearman's ρ) between the behavioral metric and citations in the following year cit^{+1} with their corresponding upper and lower bounds of the 99% confidence interval. 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).

	post		*view*		*exp*	
tag stem	Pearson's r	Spearman's ρ	Pearson's r	Spearman's ρ	Pearson's r	Spearman's ρ
model	.246 [.208, .284]	.264 [.227, .301]	.101 [.045, .156]	.184 [.129, .237]	.167 [.087, .246]	.257 [.179, .332]
web	.492 [.443, .538]	.381 [.326, .433]	.196 [.137, .253]	.293 [.236, .347]	.309 [.241, .374]	.287 [.218, .353]
inform	.357 [.300, .412]	.399 [.343, .451]	.144 [.066, .220]	.272 [.198, .344]	.510 [.430, .582]	.327 [.232, .416]
network	.425 [.379, .469]	.356 [.307, .403]	.156 [.092, .219]	.269 [.208, .328]	.232 [.151, .309]	.274 [.196, .350]
system	.387 [.335, .436]	.290 [.235, .344]	.109 [.026, .190]	.196 [.114, .275]	.159 [.046, .267]	.279 [.171, .381]
semant	.302 [.255, .348]	.306 [.259, .352]	.144 [.091, .196]	.245 [.194, .295]	.199 [.134, .263]	.209 [.144, .272]
analysi	.316 [.270, .361]	.281 [.234, .327]	.106 [.044, .168]	.248 [.188, .306]	.210 [.132, .286]	.217 [.139, .293]
ontolog	.355 [.303, .405]	.312 [.259, .364]	.151 [.096, .206]	.255 [.201, .306]	.262 [.197, .324]	.285 [.221, .346]
social	.287 [.222, .350]	.379 [.317, .437]	.068 [-.007, .142]	.238 [.166, .307]	.094 [.001, .185]	.274 [.186, .357]
commun	.257 [.185, .325]	.301 [.231, .367]	.191 [.114, .265]	.295 [.221, .364]	.428 [.339, .510]	.290 [.191, .383]
theori	.250 [.187, .312]	.203 [.138, .266]	.235 [.152, .316]	.266 [.183, .345]	.562 [.473, .639]	.256 [.139, .365]
comput	.577 [.525, .624]	.281 [.211, .348]	.394 [.306, .475]	.112 [.012, .209]	.546 [.452, .629]	.220 [.098, .336]
learn	.274 [.219, .328]	.338 [.285, .389]	.093 [.033, .153]	.175 [.116, .233]	.136 [.052, .218]	.242 [.161, .320]
manag	.168 [.074, .259]	.216 [.123, .304]	.037 [-.073, .146]	.121 [.011, .228]	.110 [-.031, .247]	.167 [.027, .300]
tag	.420 [.348, .487]	.395 [.321, .464]	.219 [.139, .296]	.301 [.224, .374]	.304 [.215, .388]	.399 [.316, .477]
design	.459 [.404, .511]	.290 [.227, .351]	.391 [.325, .454]	.146 [.071, .220]	.601 [.529, .665]	.170 [.065, .271]
evalu	.219 [.130, .305]	.229 [.140, .314]	.177 [.084, .268]	.184 [.090, .274]	.344 [.229, .449]	.224 [.103, .339]
folksonomi	.339 [.249, .423]	.414 [.329, .492]	.120 [.025, .212]	.352 [.266, .432]	.196 [.091, .297]	.376 [.281, .464]
softwar	.314 [.249, .377]	.193 [.124, .260]	.152 [.080, .223]	.207 [.136, .276]	.386 [.300, .466]	.225 [.130, .315]
collabor	.473 [.409, .533]	.432 [.365, .495]	.196 [.118, .271]	.311 [.237, .381]	.264 [.169, .353]	.324 [.233, .410]
data	.229 [.168, .289]	.222 [.160, .282]	.179 [.091, .264]	.328 [.247, .406]	.216 [.099, .327]	.315 [.203, .419]
knowledg	.204 [.120, .285]	.252 [.170, .331]	.147 [.061, .231]	.240 [.157, .321]	.190 [.076, .300]	.206 [.092, .314]
process	.238 [.178, .296]	.220 [.159, .279]	.096 [-.002, .191]	.176 [.079, .268]	.183 [.041, .317]	.130 [-.012, .268]
cluster	.344 [.279, .405]	.411 [.350, .469]	.195 [.121, .267]	.283 [.211, .351]	.277 [.180, .368]	.254 [.157, .347]
algorithm	.316 [.282, .350]	.245 [.210, .280]	.256 [.194, .317]	.278 [.217, .337]	.181 [.086, .273]	.362 [.275, .443]
web20	.416 [.325, .499]	.451 [.363, .530]	.214 [.113, .311]	.321 [.225, .411]	.327 [.216, .430]	.366 [.257, .465]
internet	.259 [.118, .390]	.224 [.081, .358]	.142 [-.020, .296]	.139 [-.023, .293]	.136 [-.067, .328]	.265 [.068, .443]
search	.355 [.278, .428]	.320 [.241, .395]	.158 [.069, .244]	.211 [.123, .295]	.251 [.145, .352]	.237 [.130, .338]
structur	.439 [.366, .506]	.400 [.324, .470]	.198 [.073, .317]	.421 [.310, .520]	.278 [.116, .425]	.473 [.332, .593]
languag	.209 [.124, .291]	.286 [.204, .364]	.108 [.012, .203]	.191 [.097, .282]	.205 [.086, .319]	.183 [.064, .298]

Table A.6: *Continued*: For each tag stem, the correlations (Pearson's r and Spearman's ρ) between the behavioral metric and citations in the following year cit^{+1} with their corresponding upper and lower bounds of the 99% confidence interval. 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).

	exp_{Bib}		*req*		*tag*	
tag stem	Pearson's r	Spearman's ρ	Pearson's r	Spearman's ρ	Pearson's r	Spearman's ρ
model	.156 [.072, .237]	.259 [.178, .336]	.118 [.065, .169]	.216 [.165, .266]	.055 [.034, .076]	.145 [.125, .165]
web	.315 [.246, .380]	.295 [.225, .362]	.231 [.176, .285]	.289 [.236, .340]	.155 [.114, .196]	.076 [.035, .117]
inform	.533 [.453, .604]	.324 [.227, .415]	.228 [.157, .296]	.294 [.226, .359]	.132 [.096, .169]	.088 [.051, .124]
network	.233 [.151, .312]	.278 [.198, .355]	.187 [.128, .245]	.259 [.202, .315]	.150 [.121, .178]	.119 [.090, .148]
system	.154 [.038, .265]	.273 [.161, .377]	.124 [.047, .199]	.199 [.124, .272]	.087 [.055, .118]	.133 [.102, .164]
semant	.186 [.120, .251]	.214 [.149, .278]	.163 [.112, .212]	.243 [.194, .291]	.149 [.115, .183]	.162 [.128, .196]
analysi	.222 [.142, .299]	.229 [.149, .306]	.140 [.082, .197]	.261 [.205, .315]	.046 [.019, .072]	.111 [.084, .137]
ontolog	.270 [.204, .333]	.281 [.215, .343]	.185 [.134, .236]	.269 [.219, .317]	.130 [.093, .168]	.130 [.092, .167]
social	.097 [.002, .190]	.274 [.184, .360]	.081 [.009, .151]	.252 [.184, .318]	.063 [.019, .107]	.109 [.065, .153]
commun	.477 [.390, .556]	.268 [.166, .365]	.260 [.189, .328]	.319 [.251, .384]	.062 [.017, .107]	.172 [.128, .216]
theori	.562 [.470, .642]	.253 [.133, .366]	.344 [.270, .414]	.282 [.205, .355]	.036 [.002, .071]	.170 [.136, .203]
comput	.524 [.423, .612]	.214 [.087, .334]	.450 [.375, .519]	.131 [.041, .218]	.056 [.019, .093]	.055 [.018, .092]
learn	.140 [.053, .225]	.271 [.187, .350]	.120 [.063, .176]	.195 [.140, .250]	-.004 [-.037, .029]	-.007 [-.040, .026]
manag	.076 [-.069, .218]	.199 [.056, .333]	.049 [-.054, .151]	.151 [.050, .250]	.027 [-.028, .081]	.071 [.017, .126]
tag	.325 [.235, .409]	.411 [.327, .488]	.241 [.164, .315]	.304 [.229, .375]	.327 [.270, .382]	.352 [.295, .406]
design	.617 [.543, .681]	.179 [.069, .285]	.471 [.414, .525]	.155 [.084, .224]	.036 [.000, .071]	.105 [.070, .140]
evalu	.342 [.225, .450]	.228 [.104, .345]	.225 [.138, .309]	.199 [.111, .284]	.049 [-.002, .099]	.083 [.033, .133]
folksonomi	.197 [.091, .299]	.379 [.282, .468]	.136 [.044, .227]	.350 [.266, .429]	.212 [.135, .287]	.338 [.266, .407]
softwar	.405 [.319, .485]	.241 [.145, .332]	.204 [.137, .269]	.181 [.113, .246]	.131 [.089, .172]	.003 [-.039, .045]
collabor	.255 [.159, .347]	.314 [.220, .402]	.223 [.150, .294]	.339 [.270, .404]	.155 [.104, .205]	.153 [.102, .203]
data	.212 [.091, .326]	.305 [.189, .412]	.237 [.158, .313]	.323 [.247, .395]	.122 [.090, .154]	.052 [.020, .085]
knowledg	.156 [.038, .270]	.222 [.106, .332]	.151 [.070, .230]	.232 [.153, .308]	.039 [-.012, .090]	.003 [-.048, .054]
process	.144 [-.003, .284]	.175 [.030, .314]	.133 [.042, .221]	.154 [.063, .242]	.078 [.046, .110]	.268 [.238, .298]
cluster	.245 [.145, .340]	.250 [.150, .345]	.232 [.162, .299]	.286 [.218, .351]	.168 [.127, .207]	.302 [.265, .339]
algorithm	.180 [.082, .274]	.362 [.273, .445]	.205 [.146, .262]	.290 [.233, .344]	.034 [.014, .054]	-.021 [-.041, -.002]
web20	.378 [.269, .477]	.389 [.281, .487]	.241 [.142, .335]	.322 [.227, .410]	.356 [.279, .429]	.289 [.208, .366]
internet	.158 [-.049, .352]	.306 [.108, .481]	.154 [.006, .295]	.142 [-.006, .284]	.122 [.045, .198]	-.017 [-.094, .061]
search	.258 [.151, .360]	.242 [.134, .345]	.082 [-.003, .166]	.204 [.121, .284]	.227 [.174, .278]	.147 [.093, .200]
structur	.259 [.093, .410]	.448 [.301, .574]	.235 [.122, .343]	.428 [.327, .519]	.259 [.217, .300]	.175 [.132, .218]
languag	.194 [.072, .310]	.195 [.074, .311]	.138 [.050, .225]	.183 [.096, .268]	.015 [-.031, .061]	.086 [.040, .132]