



Social impact assessment of scientist from mainstream news and weblogs

Title	Social impact assessment of scientist from mainstream news and weblogs
Author(s)	Timilsina, Mohan;Khawaja, Waqas;Davis, Brian;Taylor, Mike;Hayes, Conor
Publication Date	2017-10-14
Publisher	Springer Verlag
Repository DOI	10.1007/s13278-017-0466-x

Social Impact Assessment of Scientist from Mainstream News and Weblogs.

Mohan Timilsina · Waqas Khawaja ·
Brian Davis · Mike Taylor · Conor
Hayes

Abstract Research policy makers, funding agencies, universities, and government organizations evaluate research output or impact based on the traditional citation count, peer review, h-index and journal impact factors. These impact measures also known as bibliometric indicators are limited to the academic community and cannot provide the broad perspective of research impact in public, government or business. The understanding that scholarly impact outside scientific and academic sphere has given rise to an area of scientometrics called alternative metrics or "altmetrics". Moreover, researchers in this area incline to center around gauging scientific activity via social media namely Twitter. However, these count-based measurements of impact are sensitive to gaming as they lack concrete references to the primary source. In this work, we expand a conventional citation graph to a heterogeneous graph of publications, scientists, venues, organizations based on more reliable social media sources such as mainstream news and weblogs. Our method is composed of two components: the first one is combining the bibliometric data with social media data like blogs and mainstream news. The second component investigates how standard graph-based metrics can be applied to a heterogeneous graph to predict the academic impact. Our result showed moderate correlations and positive associations between the computed graph-based metrics with academic impact and also reasonably predict the academic impact of researchers.

Keywords altmetrics · heterogeneous · graph · impact · h-index · scientist · prediction

Mohan Timilsina · Waqas Khawaja · Brian Davis · Conor Hayes
Insight Centre for Data Analytics, National University of Ireland Galway
E-mail: firstname.lastname@insight-centre.org

Mike Taylor
Statistical Cybermetrics Research Group, Wolverhampton University, United Kingdom
E-mail: mike@manometrics.com

1 Introduction

This paper extends our previous work on prediction of academic impact from mainstream news and weblogs [1]. The main distinction of the method described in this paper from the preceding work can be found in three facets. The first is the how we extracted the names of the scientist mentioned in a social media and disambiguate them. The second is our experiment about predicting absolute h-index using different graph-based influence metrics. The third facet is how to categorize the scientist on the basis of social versus academic presence.

Traditional impact indicators such as citation counts, h-index and journal impact factor [2,3] are restricted to the academic community and they do not capture the wider socio-economic research impact i.e impact at the general public, government or business levels. In recent times, some academics have become increasingly dissatisfied with the use of traditional bibliometric indicators arguing that the traditional measures of scientific impact are too slow to accurately capture scientific output in the modern information age [4].

This aforementioned limitation of traditional metrics led to the expansion of novel, alternative measures of scientific impact - Altmetrics [5]. Altmetrics is the blend of **(a)** alternative data sources and **(b)** metrics derived from these sources. The combined attributes attempt to use the web as a platform from which to investigate and measure the extent to which scientific work finds its way to non-academic audiences. Therefore most of these metrics involve measuring web-based activity surrounding scientific articles, authors, and institutions.

The *Impact* of scientists in bibliometrics is traditionally measured using the **h-index** score. The **h-index** score for a scientist is defined by having **h** publications which have been cited at least **h** times. A high impact scientist is therefore highly cited by an academic community. With the advent of the Web, the discourse surrounding scientific work has moved from purely academic domains to wider areas of discourse. In this scenario, there is a need to measure the broad impact of scholarly resources outside the scientific community. The current trend of measuring the impact of scholarly activity in social media is based on a count of bookmarks, blog posts, views, tweets, likes, shares, etc. A count metric is typically considered as a measure of influence by the scientific article or scientist in social media but this is misleading because it is difficult to prove that any publication or scientist receiving 1000 *tweets* or *likes* implies that it is highly influential. The reason is counts can be gamed or the publication has catchy or funny headlines. The activity in social media like Facebook or Twitter are neutral, mere pointers to research than comments or discussion [6,7]. In order to address this apparent weakness in social media, we chose to use lengthier documents, for example, blogs and mainstream news references. We suggest that when a researcher, institute or publication work is mentioned or linked in such media, then they are more likely to be impactful in a social context, and gaining value in their field. Similarly, a mention in the non-scholarly literature like mainstream news and blogs will bring more

attention to this research output than other forms of social media [8,9]. Due to this, we propose an approach to measure the impact of scientists in the non-scholarly literature as a means to measure their social impact and to predict the academic impact.

The remainder of this article is framed as follows: In Section 2 we review related work concerning graph based influence metrics of scientific literature and scientist in citations and co-authorship networks. In Section 3 we describe our dataset and provide details of how we construct our heterogeneous graph of social media and scientists. We also implemented different centrality metrics to assess the influence of scientist in a heterogeneous network and performed the correlation significance test between the computed graph-based metrics and **h-index** of the scientist. We obtain candidate measures which are the basis for our prediction method described in Section 3. We outline and clarify our findings and discuss their implications and future work in Section 4.

2 Related Work

The traditional measure of scientific publications is based on the citation counts proposed by Garfield [10]. A higher number of citations of scientific publications garners attention in the scientific community because it indicates importance. In the context of citation graphs of scientific publications, Google's PageRank algorithm [11] deployed in a citation network brought insight to measure the research impact of scientific publications. Bollen et al. [12] implemented the PageRank algorithm to rank the scientific publications in a temporal network. The algorithm is also applied on the co-author network [13,14] to rank the influential scientist. **Article Influence Score** [15,16], a metric inspired from PageRank to measure journals total importance to the scientific community.

The graph theory approach provides a solid foundation for ranking scientific publication and scientists in the context of a homogeneous network for example network of citations between publications. In the case of ranking network entities in a heterogeneous network, these metrics are not useful. Zhou et al. [17] came up with the heterogeneous network approach to computing the impact of researchers and publications using different kinds of networks, for example, the social network of authors, citation network of publications and authorship network connecting the publications and authors network. This model provides a **co-ranking** of articles and authors. The problem with **co-ranking** model was it ranked publications based on its previous popularity, so the recent publications always receive lower scores and thus it was not useful to predict the influence of latest publications. Sayyadi [18] proposed a FutureRank algorithm which combined the information about citations, publication time and authors to rank the scientific articles by predicting the future ranking. Both Zhou and Sayyadi [18,17] did not clarify whether their method can be extended to rank other academic entities such as institutions and scientific venues. Their methods were limited to the only citation, co-authorship, and

author network did not clearly mention whether such network can integrate with other kinds of the network for example academic organizational network and measure their impact.

Sarigol et al. [19] studied the centrality of scientific authors in a co-authorship network and use the computed metrics to predict the citations of their publications. Their study focused solely on the computer science research domain, hence we must take into account that a co-authorship network may vary from discipline to discipline. Furthermore, scientific reputation plays an important role for the scientist in his or her publication's citation rate.

Social media has provided an instantaneous means to disseminate scientific work and has enabled researchers to contribute to the building of research communities [20,21]. Li et al. [22] explored the measure of scholars influence in academic social media platforms, considering both the academic and social impact. **Mendeley**¹ data was used and network centrality metrics were applied to measure the social influence. They reported that those scholars with high academic impact are not necessarily influential from the social point of view. This study was only conducted in Mendeley data so it might not comprehensively reflect scholars influence only accounting single social media [23].

In the context of academic social media Hoffmann et al. [24] introduced Impact Factor 2.0 to measure an impact of researchers. The social network of Swiss management scholars on **ResearchGate**² was analyzed using network centrality measures. They reported significant correlations between computed social network metrics such as *eigenvector centrality*, *indegree*, and *closeness centrality* with h-index a traditional bibliometric measure. The caveat of their study is the small sample size of only 45 researchers and that the data was only sourced from the ResearchGate. The findings of their studies could be biased because researchers use multiple social networks [25] such as Twitter, Facebook, Mendeley, Blogs, etc.

Acuna et al. [26] attempted to predict h-index of neuroscientists from the features extracted from their CV using regression equations. The important finding of their approach is the academic CV, the reason for that feature extracted from CV can be used as alternative data source to predict the impact of neuroscientists. Ringelhan et al. [27] studied unpublished scientific articles receiving likes in Facebook as an early indicator to predict the impact of scientific work. A common issue with using social networks is that the scientific community may not consider Facebook *likes*, Twitter *tweets/retweets* as legitimate sources because they can be manipulated or gamed [28].

Most of the prediction analysis have been performed on the bibliometric data sets [26,29,30] but few of the initiative were taken to predict the scientific impact using social media data [31,27]. In this work, however, we focus on blogs and mainstream news because these media bring attention to research output than any other social media [32]. To the best of our knowledge, there is no such integrated heterogeneous graph-based approach between bibliometrics

¹ <https://www.mendeley.com/>

² <http://www.researchgate.net/>

data with social data such as blogs, mainstream news to measure and predict the academic impacts.

3 Methodology

We investigated three research questions, aiming to measure the academic impacts in social media like blogs and mainstream news. First, we examined whether we can integrate bibliometric data to social media data to create a heterogeneous network. Second, we investigated the centrality metrics of a scientist in such network. Thirdly and finally, we explored how centrality metrics can predict the academic impact. We start with describing the first research question.

Can we integrate the blogs and mainstream news with Bibliometric data?

To answer the first research question we performed the following steps:

1. **Collection of Data:** We began with the social media data. Our data is *Spinn3r*³ data which is a crawl of the blogosphere for the time period of 2010 November to 2011 July. The data is stored in a distributed file system and has eight publisher types: *memetracker*, *forum*, *microblog*, *review*, *classified*, *mainstream news*, *weblog* and *social media*. We extracted only weblogs and mainstream news from this distributed file using Java Spinn3r API⁴ and the collected data are stored in a MongoDB⁵ database which stores the data as JSON⁶ documents. We indexed extracted data using Solr⁷ for quick search of the topic of interest. For the bibliometric data, we used SCOPUS⁸, one of the largest bibliographic database which contains the citations of peer-reviewed literature: scientific journals, books and conference proceedings. We used the Elsevier SCOPUS API⁹ to extract metadata of publications such as citations, authors, publication venue and organizations. The extracted data are in JSON format and are indexed.
2. **Search of a Candidate Topic:** In order to find out the connectivity between the two types of data sources, we restricted our focus on a topic that has received a lot of public attention in the time window of our social media index (Nov 2010-July 2011). We used Wikipedia¹⁰ to research

³ <http://spinn3r.com/>

⁴ <http://www.programmableweb.com/api/spinn3r>

⁵ <https://www.mongodb.org/>

⁶ <http://www.json.org/>

⁷ <http://lucene.apache.org/solr/>

⁸ <https://www.scopus.com/home.uri>

⁹ <http://dev.elsevier.com>

¹⁰ <https://www.wikipedia.org/>

prominent news events recorded in that period. This suggested one public health topic was particularly newsworthy: The emergence of a virulent strain of *Avian Influenza*. An examination of query trends in the Google search engine suggests bursts in Web user interest in these topics in the analysis period as shown in figure 1.



Fig. 1 Google Trend for the Query 'Avian Influenza' ; from Nov 2010 - July 2011

We created a focused subset of the data by extracting from the Spinn3r and SCOPUS data sources only the content related to our focus topic. To do so, we issued queries over our collections and extracted the content items mentioning the synonymous phrases that all refer to avian flu: "bird flu", "avian influenza", "H5N1", "avian flu", "fowl plague", "grippe aviaire". This dataset restriction has brought our experimental data to a manageable size, making it ideal for preliminary analysis and experiments. We collected 259,149 JSON documents from Spinn3r dataset and 37,081 scientific publications from SCOPUS dataset.

3. **Construction of Graph Data Model:** We took the same graph data model from our previous work [1]. This model used the conceptual model of graph data from the system architecture of Targeted Elsevier Project at **Insight Centre for Data Analytics**¹¹ which consists of seven different types of node entities and five different types of relationships entities. The figure 2 shows the graph data model used for storing the data and for analysis: The definition of each node and relationship is shown in Table 1 and 2 respectively.

¹¹ <http://www.insight-centre.org/>

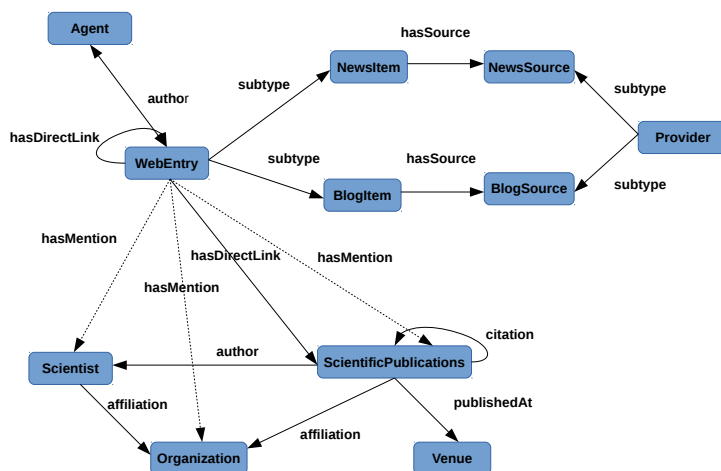


Fig. 2 Conceptual Graph Data Model

Relations Type	Definition
<i>hasDirectLink</i>	This relationship occurs between web entries or between web entries and a publication. These relation are directly extracted from HTML content of web entries as anchors.
<i>hasMention</i>	This relationship is not directly extracted from the data. The relationships are extracted using the text analysis methods like entity extraction, disambiguation, and linking.
<i>hasSource</i>	This relationship occurs between the web entries and its source.
<i>Citation</i>	The citation relationship connects two scientific publications. This relationship is extracted using SCOPUS API.
<i>Author and Affiliation</i>	These relationships between author, publications, and authors are extracted from SCOPUS API.

Table 2 Description of the Relationships in the Graph Model

In order to store the information of nodes and relationships we used the **Neo4J**¹² graph database. Neo4J was chosen as it is a free and open source graph database and has APIs for most of the popular program-

¹² <http://neo4j.com/>

Node Type	Definition
<i>Web Entry</i>	The Web Entry are the nodes correspond to the items on the web. The items here are the Blogs and Mainstream News which were extracted from the spinn3r data. Each of these particular types of entries have a corresponding node type as subtype for the WebResource. Properties: url, full text, timestamp
<i>Agent</i>	The agent nodes are the authors of the web entries. The author is the username of the account who produced the entry. Properties: username, email, homepage
<i>Provider</i>	The provider nodes are the sources of the web data, for example: www.theguardian.com , www.twitter.com etc. The provider node has the subtype for example NewsSource, BlogSource. Each web resources are linked by hasSource link to the corresponding provider. Properties: url
<i>ScientificPublication</i>	The <i>ScientificPublication</i> type corresponds to nodes that are as the same name as the scientific publications. Properties: url, text, abstract
<i>Scientist</i>	Scientist types nodes are the authors of the scientific publications. Properties: name, email
<i>Organization</i>	Organization type nodes are the universities or the research institutions which are extracted from the author’s affiliations. Properties: name, website
<i>Venue</i>	The venue type nodes represent journal, conferences, workshops, etc. Properties: name, website

Table 1 Description of the Nodes in the Graph Model

ming languages like Java, Python, etc. We used the Neo4j Python API called py2neo¹³ to construct the graph.

4. Data Integration and Scientist Identification:

To identify the mentions of scientists within Spinn3r content data, we took a hybrid knowledge/learning based approach by combining an existing supervised approach with handcrafted extraction rules at the post-processing stage. We then developed a pipeline using General Architecture for Text Engineering (GATE) [33] that used a combination of ANNIE Named Entity Recogniser (NER) [34] and Stanford NER[35] to identify person names. We crafted custom JAPE grammar rules to annotate mentions prioritizing certain ANNIE annotations over Stanford NER annotations for Person Names containing punctuation (i.e *Dr J. Smith*) as these were problematic for the Stanford classifier. In addition, the extract rules took advantage of additional linguistic context such as whether the mention of the scientist

¹³ <http://py2neo.org/2.0/>

was contained in a quotation i.e. *Dr M. Knight says "...*". These preferences were set according to the manual observation of results from these two.

JAPE (Java Annotations Pattern Engine) is a pattern matching language over features and annotations implemented as a cascade of finite-state transducers [36]. We ran our pipeline [37] over the contents of Spinn3r data and identified person names within our corpus.

We prepared a list of scientist names in parallel from SCOPUS and indexed them using Lucene¹⁴. SCOPUS provided multiple possible variants of how scientist names are mentioned in the literature. We then used MongeAlkan [38] string similarity to match the person names identified from our pipeline to scientist names indexed from SCOPUS using a threshold of 0.99 after observing results from a few string comparison methods as shown in Table 3. We were able to identify almost 2351 scientist names.

In order to avert the disambiguation and linking the problem of differentiating between multiple scientists with the same name, we inspected only the research profiles with unique surname and name combination similar to the study done by [39,40]. We then checked manually those names who actually published papers related to "Avian Influenza". Hence, we are left with a relatively small subsets of **320** scientists within a specific topic and are free from the name disambiguation problem. The next step is to link the identified name of scientists from Spinn3r to the SCOPUS graph. The overall process is shown in the Figure 3.

¹⁴ <http://lucene.apache.org/core/>

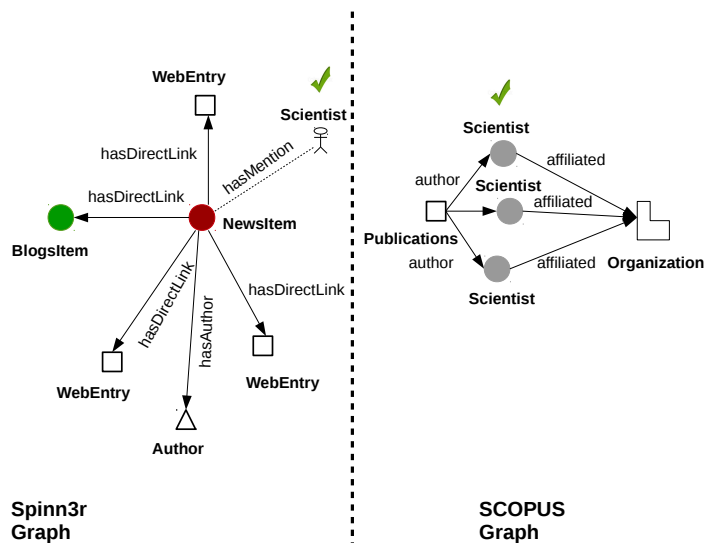


Fig. 3 Integration between Spinn3r and SCOPUS Graph Data

In Figure 3, the **NewsItem** is a node in a Spinn3r graph where a scientist is mentioned shown by a dotted line. The information about the identified scientist is in the SCOPUS graph. The **Scientist** is a node in a SCOPUS graph. We linked the scientists who are identified in a Spinn3r graph to SCOPUS Graph using **hasMention** relationship. In order to carry out this procedure, we issued a Neo4J Cypher¹⁵ query which connects the Spinn3r graph web entries nodes to a SCOPUS graph scientist node through **hasMention** relationships. Consequently, we were able to link **320** scientists in our SCOPUS graph.

5. **Graph Dataset Statistics** The different types of nodes and relationship count is shown in Table 4 and 5 in a connected Spinn3r/SCOPUS graph.

¹⁵ <https://neo4j.com/developer/cypher-query-language/>

Name of the Scientist		Similarity Score		
		MongeElkan	Cosine	Levenshtein
Johnson	Johnson	1	1	
Johnson	Avery Johnson	1	0.7	
Johnson	Don Graham	0.6		
Johnson	Joe Sakic	0.6		
Johnson	Melanson	0.5		
Wade	Wade	1	1	1
Wade	Bill Walton	0.5		
Wade	Walton	0.5		
Wade	Dwayne Wade	1	0.7	
Wade	Pat	0.5		
Wade	Sam Carchidi	0.5		
Wade	Ryan Wittman	0.5		

Table 3 Comparing Scientist Names

Nodes	Count
Mainstream News	10035
Weblogs	79268
News Sources	1717
Blog Sources	11699
Web Entry	828311
Scientist	320

Table 4 Node Types With their Count

Relationship	Count
hasDirectLinks	5408825
hasAuthor (of web content)	89978
publishedAt	16584
hasSource	95275
author (of scientific publication)	99986
hasMention	320
affiliation	77234

Table 5 Relationship Types With their Count

Finally, we constructed a graph with integrated the bibliometrics and social media data using *hasMention* relationships. In this process, we linked **320** scientists mentioned in social media. In the next section, we will address the second research question about measuring the importance of those scientist mentioned in social media.

Can we measure the influence of scientists who are mentioned in blogs and mainstream news?

In order to answer this research question, we used the following methods:

1. **Mention Count:** The mention count of a scientist is the number of times the scientist was mentioned in social media. In other words, mention count is the *indegree* of the scientist node in a bipartite graph between the Web entry and the scientist node with **hasMention** relationship.
2. **PageRank Score:** We computed the PageRank [41] score of all the blogs and mainstream news nodes in a hyperlink network. We summed all the PageRank of those web entry where the scientist mentioned. Thus the impact of scientist based on PageRank Score is given by:

$$Scientist_{(influence_{PR})} = \sum_{i=1}^n WebEntry(PR_i) \quad (1)$$

PR is the PageRank score of the web entry node and n is the total number of web entry where the scientist is mentioned.

Figure 4 shows example of the higher PageRank and the lower PageRank impact of scientist mentioned in social media.

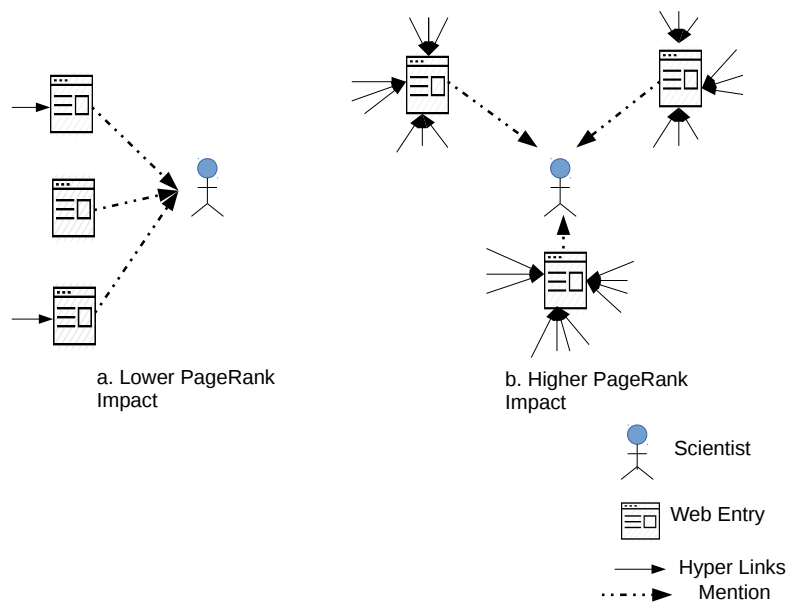


Fig. 4 PageRank Score of Scientist Mention in Social Media.

3. **Authority Score:** We computed the Authority Score using the HITS (hyperlink-induced topic search) authority algorithm [42] of all blogs and mainstream news in a hyperlink network. We summed all the Authority Score of those web entry where the scientist mentioned. Thus the impact of scientist based on Authority Score is given by:

$$Scientist_{(influence_A)} = \sum_{i=1}^n WebEntry(A_i) \quad (2)$$

A is the Authority score of the web entry node and n is the total number of web entry where scientist is mentioned.

Figure 5 shows an example of the higher authority and lower authority scores impact of scientist mentioned in a social media.

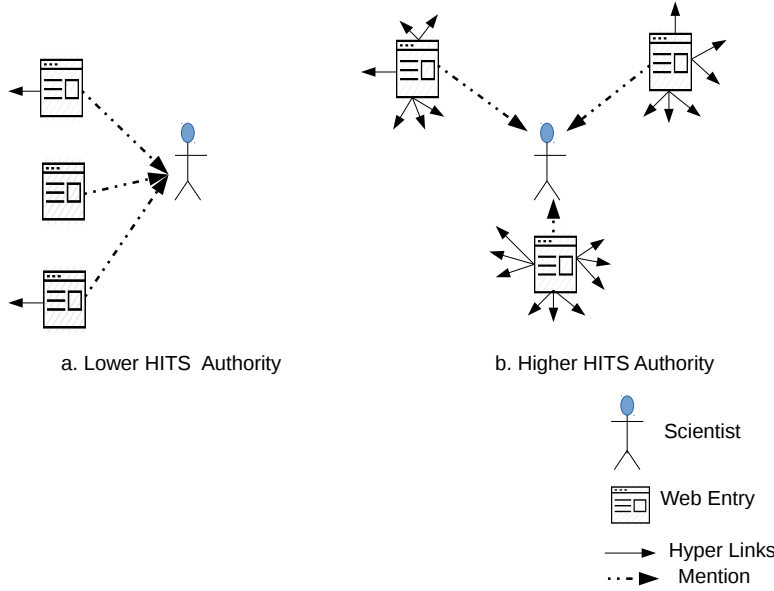


Fig. 5 Authority Score of Scientist Mention in Social Media.

4. **Unweighted Node Count:** It is the total count of the directly and indirectly linked nodes in a *maximal directed ego subgraph* where the root node is the scientist and the other nodes are the web entries referred through the hyperlink relationships. The definition of *Directed Ego-Centered Graph* and *Maximal Directed Ego Network* is given as:

Definition 1: Directed Ego Centered Graph: For a graph $G = (V, E)$ where V is the set of nodes and $E \subseteq V \times V$ is a set of ordered pairs from

V called the edges of the graph, the ego network of k^{th} degree is given by $G_i^k = (s_i \cup V_i^k, E_i)$ where V_i^k is the set of nodes that are at most k hops away from s_i and E_i is the set of directed edges between $s_i \cup V_i^k$ and s_i the seed node of graph G_i^k .

Definition 2: Maximal Directed Ego Network: A maximal directed ego network of a graph $G = (V, E)$ is an ego network of k hop away from the node s_i given by $G_i^k = (s_i \cup V_i^k, E_i)$ such that there is no vertex in $V \setminus V_i^k$ whose addition in G_i^k would preserve the property of a directed ego centered network.

Fig 6 shows the example of Unweighted Node Count in a Maximal Directed Ego Network.

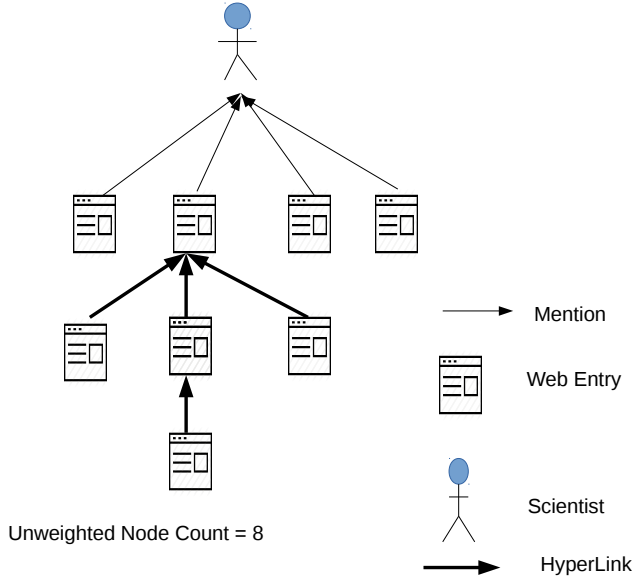


Fig. 6 Unweighted Node Count in a Maximal Directed Ego Network of Scientist

5. **Katz Centrality:** The Katz centrality [43] is applied to a maximal directed ego network of the web entry nodes in a hyperlink network where the scientist is mentioned. The combined score based on Katz Centrality score is given by:

$$Scientist_{(influence_K)} = \sum_{i=1}^n WebEntry(K_i) \quad (3)$$

The K is computed as follows:

$$K = \sum_{j=1}^{\infty} \sum_{i=1}^d \alpha^j (A^j)_{ij} \quad (4)$$

K is the Katz Centrality of the web entry node in a maximal directed network and n is the total number of web entries where the scientist is mentioned. A is the adjacency matrix of the graph, α is the reciprocal of the eigenvalues of adjacency matrix A , d is the degree between node i and node j .

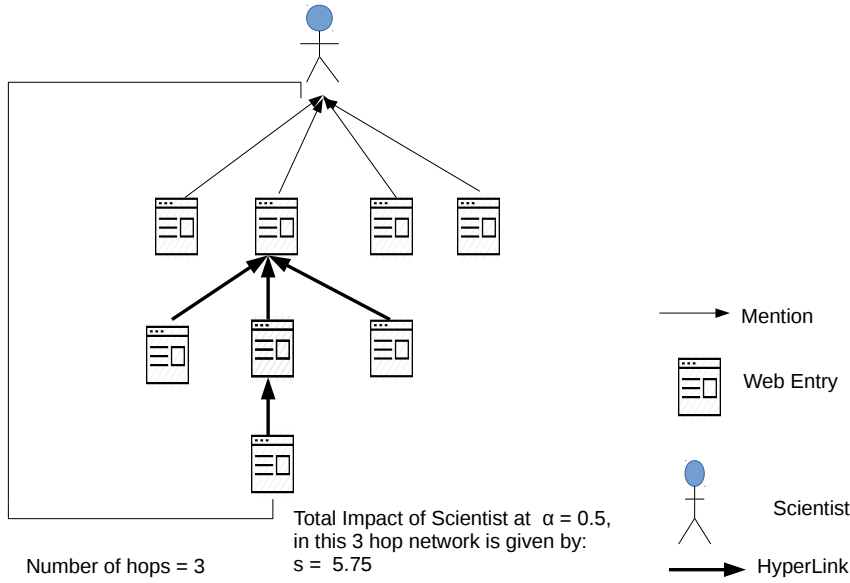


Fig. 7 Katz Centrality Score Computation based upon Scientist Mention in Social Media.

Figure 7 demonstrates the computation of the Katz Centrality score of Scientist in a 3 hop network at attenuation parameter $\alpha = 0.5$

6. **Log Based Weight:** This is the metric we propose to weight the nodes in a maximal directed ego network. Log Based Weight is based on the information spreading ability of each node. If a scientist is mentioned in a subgraph of Hyperlink network then total influence of the scientist in that subgraph based on **Log Based Weight (lbw)** is the cumulative sum of spreading ability of each node which is given by,

$$Scientist_{(influence_{lbw})} = \sum_{i=1}^N \log \left[\frac{Indegree + 1}{Outdegree + 1} + 1 \right] \quad (5)$$

The rationale to use *log* is that for a very high indegree of the web entry nodes, the score will also be very high, so we dampened the score using logarithm, and to smooth the equation for becoming unstable we added 1. Figure 8 shows the computation of Log Based Weight in three different network configurations. With respect to the first configurations, in Figure 8(a) there is a direct mention link of scientist and for the second, Figure 8(b) indicates a mention along with indirect hyperlink. With respect to the third and final configuration in Figure 8(c) there is a direct mention and a hyperlink relationship together.

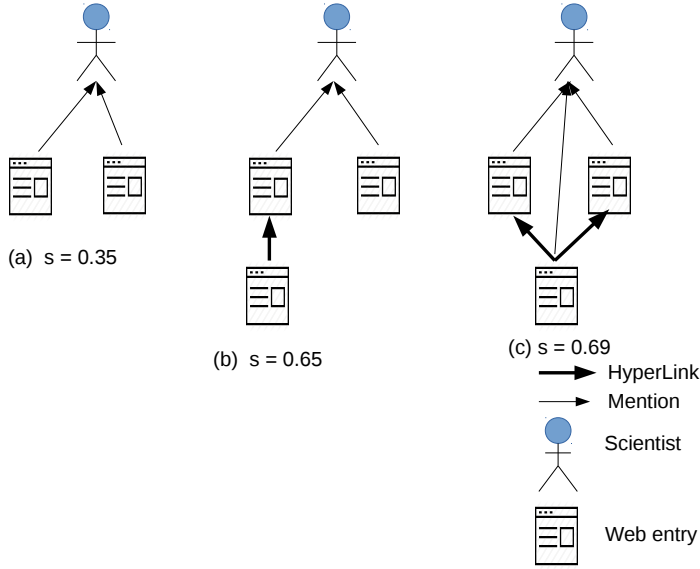


Fig. 8 Log Based Weight for 3 network configurations. The score of Scientist grows from (a) to (c)

Comparison of Different Metric with h-index for Scientist Mentioned in Social Media

We applied the metrics described above and computed the scores for 320 scientists. We performed a Spearman correlation [44] test between the computed metrics and the corresponding **h-index** of the scientist. The result of the correlation significance test is shown in the Table 6.

The computed metrics are weakly correlated but statistically significant with respect to its h-index. The significant correlation infers that there is also a correlation in the population of a scientist with their social media score. This

Metrics	Correlations	p-value $\alpha = 0.05$
Mention Count	0.35***	1.09e-10
Unweighted Node Count	0.38***	1.38e-14
PageRank Score	0.34***	3.85e-10
Authority Score	0.29***	6.44e-08
Katz Centrality	0.42***	1.4e-14
Log Based Weight	0.45***	2.2e-16

Table 6 Correlation Significance Test of the Computed Metrics with h-index

N= 320, Spearman correlation is displayed. *** $p < 0.05$

concludes to the fact that correlation from the sample of 320 scientist is not due to any random effect. Our result supports the similar kind of claim by previous studies [45–48] that citations and altmetrics are positive but weakly correlated. In comparison with other computed graph-based metrics in table 6, we observed Log Based Weight ($\rho = 0.45$, p-value = 2.2e-16) and Katz Centrality ($\rho = 0.42$, p-value = 1.4e-14). Both have slightly better correlation in magnitude with the h-index. We performed pairwise correlation test between Log Based Weight and Katz Centrality to compare their significance in correlation.

This is a case of overlapping correlation problem because we compare both the metrics with the h-index. We observed that the Log Based Weight with h-index ($\rho = 0.45$) and Katz Centrality with h-index ($\rho = 0.41$) have high correlation between Log Based Weight and Katz Centrality ($\rho = 0.90$). We formulate the following hypothesis:

H_0 **Null Hypothesis:** There is no significant correlation difference between Log based Weight and Katz Centrality with h-index

$$H_0: \rho_{lbw} = \rho_{kc}$$

H_a **Alternative Hypothesis:** The correlation measured from Log Based Weight is greater than Katz Centrality with h-index.

$$H_a: \rho_{lbw} > \rho_{kc}$$

Where ρ_{lbw} is the correlation coefficient of Log Based Weight and ρ_{kc} correlation coefficient of Katz Centrality.

We performed the test proposed by Steiger [49] called Steiger’s Z-test which computes the statistical comparisons between correlation coefficients computed of the same populations. This test is implemented in the comparing correlation **cocor**¹⁶ package in the R statistical programming language.

The computed one-tailed test indicated that the $p - value < 0.05$, which means the test fails to accept the null hypothesis and accept the alternative hypothesis that the correlation measured from Log Based Weight is statistically significantly greater than the correlation measured from the Katz Centrality.

In the next step, we tried to answer our third research question, which is to evaluate computed graph-based metrics by predicting h-index.

¹⁶ <http://comparingcorrelations.org/>

Sample size	z-score	p-value ($\alpha = 0.05$)
320	1.77	*0.03

Table 7 Correlation Significance Test between Log Based Weight and Katz Centrality

Do the computed graph-based metrics predict academic impact?

To answer this research question we started with the following steps:

3.1 Building a Prediction Model

In the previous section, we discussed how to measure the impact of a scientist in social media using graph-based metrics. In this section, we will examine how these metrics can be used to predict the impact of the scientist in the academic world. In this respect, we performed two experiments, (i) one to predict the absolute h-index of a scientist taking the graph-based metrics as a predictor variable and (ii) and experiment to classify a scientist in different categories such as **low cited**, **moderately cited**, **highly cited** and **very highly cited**.

3.1.1 Regression Model with Single Predictor

We performed the correlation among all the computed graph-based metrics of a scientist against their h-index. The result showed that Log based weight has a high Spearman correlation of $\rho = 0.45$ with h-index in comparison to other graph-based metrics 6. We used this predictor variable to predict the h-index. The model can be viewed as:

$$h\text{-index} = \beta * \text{Log_Based_Weight} + \epsilon$$

where β is the regression coefficients and ϵ is the error term while predicting the dependent variable.

The descriptive statistics are shown in Table 8:

	Min	Max	M	SD	(2)
(1) Log Based Weight	0.00041	166.4	6.658	17.74	0.45***
(2) h.index	0	41	5.96	5.52	

Table 8 Descriptive Statistics and Correlations for Single Predictor

N= 320 ; Min = Minimum; Max = Maximum; M= Mean; SD= Standard Deviation.
Spearman's correlation is displayed; *** $p < 0.05$ (two-tailed).

Regression Analysis: As shown in Figure 9 for a unit change in h-index there is a 0.12 unit change in the Log Based Weight. Log Based weight positively predicts the h-index ($\beta = .12$, $p < 0.05$). β is the regression co-efficient and its value is positive and significant.

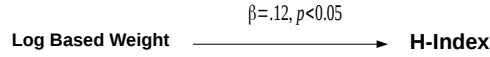


Fig. 9 Relation between Log Based Weight and h-index

3.1.2 Prediction Accuracy of Model

We performed **Leave One Out Cross Validation (LOOCV)** to check the prediction error of the model [50]. This method is known as an exhaustive cross validation method because it takes n-1 sets as a training set and performs the prediction in a single test set. The computed Root Mean Square Error (RMSE) of the model is **5.2**. The RMSE is high so this model is not highly dependable. In the next step, we perform the **Principal Component Analysis** over all the computed graph-based metrics because these metrics are highly non-independent.

3.1.3 Principal Component Analysis (PCA)

PCA converts the variable into linearly uncorrelated variables called principal components. These components capture the highest variability in the data and are known as **eigenvectors** which can be used to predict the outcome variable [51]. We applied PCA in our graph-based metrics and we found 7 different components.

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7
Standard Deviation	2.3816448	0.8488957	0.60548089	0.35624607	0.27001367	0.199283595	0.0316927959
Proportion of Variance	0.8103189	0.1029463	0.05237244	0.01813018	0.01041534	0.005673422	0.0001434905
Cumulative Proportion	0.8103189	0.9132651	0.96563757	0.98376775	0.99418309	0.999856510	1.0000000000

Table 9 Summary of Principal Component Analysis

From table 9, it is shown that both Component 1 and Component 2 capture the 91 % of the variance and other component does not contribute as much variance. Similarly, the variance contribution from Component 3 onwards is relatively small and capture a small proportion of variability and are unimportant. We choose Component 1 and Component 2 and regress with the dependent variable h-index.

We validate the model using Leave One Out Cross Validation (LOOCV). The Root Mean Squared Error (RMSE) of the model is **4.19**. The RMSE of both models is shown in Table 10:

Model	Root Mean Squared Error (RMSE)
Linear Regression	5.2
Principal Component Regression	4.19

Table 10 *RMSE Results of the Models*

We observed from Table 10 that **RMSE** with the single predictor is **5.2** and with principal component predictor is **4.19**. RMSE only reduced slightly from 5.2 to 4.19 but there is not so much difference in prediction accuracy of the model. One of the reasons for this might be the nature of the dependent variable **h-index**. The higher **h-index** gets, the harder it is to increase [52]. This means even the graph-based influence score is higher, but the h-index is not increasing. In our next experiment, we try to predict the label of the scientists which are divided into different categories according to their h-index.

3.1.4 Classifying Scientists By Their Current Social Presence

There can be four different possible combinations between the social and the academic world for any scientist. Each of the possible combinations is shown in Table 11.

Social World	Academic World
+	+
+	-
-	+
-	-

Table 11 *Social Vs Academic World: + : Active , - : Passive*

The scientists with (+,+) patterns are those who are active in the social and academic world and (+,-) patterns are those who are active in the social world, but passive in the academic world. Similarly (-,+) patterns are those who are passive in the social world but active in the academic world and (-,-) patterns are those who are passive in both social and academic worlds. In our classification problem, we tried to predict which combinations are best supported.

We used five different features from the maximal directed ego network namely, depth of the graph, number of nodes, cosine similarity between citing and cited documents, number of mentions, Log Based Weights of a scientist. The outcome variable is the category of the scientist. We categorized h-index into *four* categories using quartile distribution as shown in Table 12. This is a supervised machine learning classification problem and we trained the model using a **Support Vector Machine (SVM)**.

3.1.5 Categorization of Scientists Using h-index

The h-index of 320 scientists are divided using quartile distribution. We used each quartile as category, as seen in Table 12

Quartile Distribution of h-index	Category
(0-25)%	Low Cited
(25-50)%	Moderately Cited
(50-75)%	Highly Cited
75% above	Very Highly Cited

Table 12 Classification of Scientist according to Quartile Distribution of h-index

3.1.6 Data Splitting and Training the model

We split the data into training and test set. We took 75% data as training and 25 % data as the test set. SVM classification with the radial kernel is applied on the training data because our data was not linearly separable. We tuned the SVM parameter γ and C using 10 fold cross validation.

3.1.7 Prediction Accuracy of the Model

We compute the Precision, Recall and F1 score for each of the four classes. The precision of the model is higher for the class **Very Highly Cited** at 0.66 and lower for the class **Highly Cited** as 0.22. Similarly, the precision in predicting **Low Cited** and **Moderately Cited** class is 0.40 and 0.30 respectively.

The recall of the model is higher for predicting **Moderately Cited** class at 0.65 and lower for **Very Highly Cited** class as 0.10. Furthermore, recall for **Low Cited** class is 0.18 and **Moderately Cited** class is 0.19.

The model has a high F1 score of 0.33 for predicting **Highly cited** class and low of 0.17 for **Very Highly Cited** class. Similarly, for **Low Cited** and **Moderately Cited** class the F1 score of the model is 0.25 and 0.23 respectively. The comparison is presented in Figure 10.

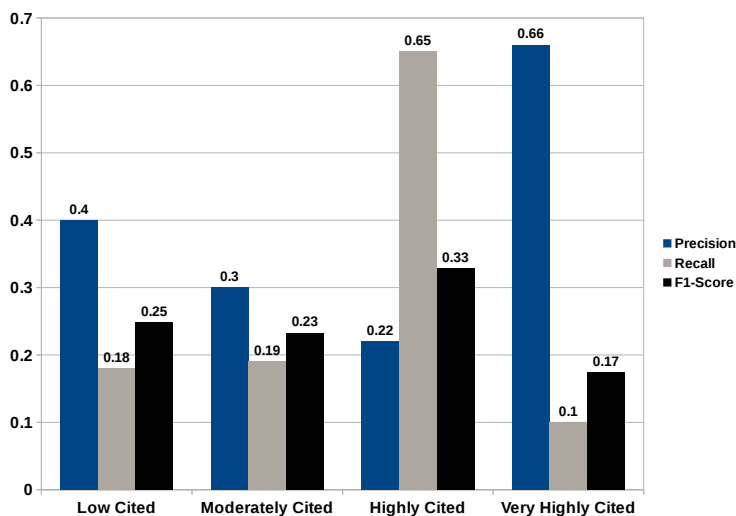


Fig. 10 Classification Accuracy of the Model in the Test Set

Discussion

The following observations are presented in Table 13.

Social World	Academic World	Possibilities Supported By Classification Algorithm
+	+	✓
+	-	X
-	+	X
-	-	X

Table 13 Combination Supported by Classification using Social media Features

The algorithm with 66 % precision and 10 % recall classifies the scientist in the category of **Very Highly Cited** class and with 22 % precision and 65 % recall classifies the scientist in the category of **Highly Cited** class. Both of these classes are above the median value of h-index in our dataset. This means our model satisfactorily classifies the scientist who is active in both social and academic world and supports the (+,+) combination. Similarly, for predicting the rest of the class the algorithm has precision and recall less than 50 %. This means algorithm cannot convincingly classify the rest of the class.

4 Limitations

In this study, we measured the impact of the scientist who is mentioned in social media. From the result of our predictive modeling, we noticed poor *F-score* and *RMSE* measures. One reason for this might be the quality of the data. In this sample, we can assume the bias towards that scientist who is both visible in social media and academics. Not all the scientists are frequently mentioned in social media platforms. In that case, it is difficult to predict the academic impact of the scientist by only taking into social media features. This may be in the case with respect to false positives (high social media presence, low academic impact) and false negative (high academic impact, low social media presence). In the experiment presented, we only use social media feature, but including features related to academia such as a number of co-authors of a scientist, number of publication in top venues or the scientist affiliation would have improved the performance ability of the classifier which we left as our future work.

Similarly, in our study, we presented the graph-based metric called **Log Based Weight**. Currently, this metric measures the information spreading ability of each node in the maximal directed ego network of a scientist. In the case of the nodes that are one hop farther from scientist node, one can assume to have a direct impact on it. While in the case of nodes which are k hops far away from scientist node, it is hard to infer the same level of influence. In our future work, we wanted to extend this metric to capture this effect.

Finally, we plan to extend our graph based centrality metrics to compute the other academic entities such as publications, venues and organizations/institutions from the heterogeneous graph. With the computed centrality metrics, we want to evaluate the metrics by predicting academic impact for other academic entities like citations for publications, impact factor for venues and mean citations scores for an organization.

5 Conclusion

In conclusion, we addressed each of our research goals as described in Section 3, by extending conventional citation graphs to heterogeneous graphs of different entities such as scientists, weblogs, and mainstream news. On a graph level, we integrate the social media data with the bibliometric data. We applied the standard graph based centrality metrics to understand the influence of scientist mentioned in social media and later we use the computed centrality metrics and their maximal directed ego network to predict the impact of a scientist. Our work extends the current trend of *Altmetrics*, which studies and seeks to measure academic impact outside from nontraditional bibliometric sources of interest, by pushing the metric boundaries beyond mere count based metrics. We achieved this by providing standard graph-based metrics for scientists which demonstrate comparable results to existing count based approaches and

demonstrate positive associations and moderate correlations to the standard bibliometric measures (h-index).

Acknowledgements We would like to acknowledge Science Foundation of Ireland (SFI/12/RC/2289) and the targeted project Elsevier for funding this research. We extend our gratitude to John Lonican for creating a citation graph from SCOPUS database and Erik Aumayr for insightful thoughts and constructive criticism. We would like to appreciate Prof. Jonice Oliveira from the Federal University of Rio de Janeiro for creative feedback and support.

References

1. M. Timilsina, B. Davis, M. Taylor, C. Hayes, in *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on* (IEEE, 2016), pp. 1388–1389
2. H.F. Moed, *Citation analysis in research evaluation*, vol. 9 (Springer Science & Business Media, 2006)
3. M. Thelwall, *Journal of information science* (2008)
4. P. McFedries, *IEEE Spectrum* **8**(49), 28 (2012)
5. C. Neylon, S. Wu, *PLoS Biol* **7**(11), e1000242 (2009)
6. M. Taylor. The challenges of measuring social impact using altmetrics [internet]. *research trends* 2013 jun [cited 2014 feb 19]; 33: 11-15 (2013)
7. D. Colquhoun, A. Plested, (2014)
8. T. Evans, (2015)
9. R. Kwok, *Nature* **500**(7463), 491 (2013)
10. E. Garfield, et al., (American Association for the Advancement of Science, 1972)
11. S. Brin, L. Page, *Computer networks* **56**(18), 3825 (2012)
12. Y. Wang, Y. Tong, M. Zeng, in *Twenty-Seventh AAAI Conference on Artificial Intelligence* (2013)
13. Y. Ding, E. Yan, A. Frazho, J. Caverlee, *Journal of the American Society for Information Science and Technology* **60**(11), 2229 (2009)
14. X. Liu, J. Bollen, M.L. Nelson, H. Van de Sompel, *Information processing & management* **41**(6), 1462 (2005)
15. C.T. Bergstrom, J.D. West, M.A. Wiseman, *The Journal of Neuroscience* **28**(45), 11433 (2008)
16. C. Bergstrom, *College & Research Libraries News* **68**(5), 314 (2007)
17. D. Zhou, S.A. Orshanskiy, H. Zha, C.L. Giles, in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on* (IEEE, 2007), pp. 739–744
18. H. Sayyadi, L. Getoor, in *SDM* (SIAM, 2009), pp. 533–544
19. E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, F. Schweitzer, *EPJ Data Science* **3**(1), 1 (2014)
20. M.V. Soto, J.E. Balls-Berry, S.G. Bishop, L.A. Aase, F.K. Timimi, V.M. Montori, C.A. Patten, *JMIR Research Protocols* **5**(3) (2016)
21. K. O'Brien, in *Seminars in Orthodontics* (Elsevier, 2016)
22. N. Li, D. Gillet, in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ACM, 2013), pp. 608–614
23. E. Mohammadi, M. Thelwall, S. Haustein, V. Larivière, *Journal of the Association for Information Science and Technology* **66**(9), 1832 (2015)
24. C.P. Hoffmann, C. Lutz, M. Meckel, in *System Sciences (HICSS), 2014 47th Hawaii International Conference on* (IEEE, 2014), pp. 1576–1585
25. A. Gruzd, M. Goertzen, in *System Sciences (HICSS), 2013 46th Hawaii International Conference on* (IEEE, 2013), pp. 3332–3341
26. D.E. Acuna, S. Allesina, K.P. Kording, *Nature* **489**(7415), 201 (2012)
27. S. Ringelhan, J. Wollersheim, I.M. Welpé, *PloS one* **10**(8), e0134389 (2015)
28. B. Hammarfelt, S. de Rijcke, A.D. Rushforth, *Information Research* **21**(2), 21 (2016)

29. A. Mazloumian, PloS one **7**(11), e49246 (2012)
30. X. Zhu, P. Turney, D. Lemire, A. Vellino, Journal of the Association for Information Science and Technology **66**(2), 408 (2015)
31. G. Eysenbach, Journal of medical Internet research **13**(4), e123 (2011)
32. A. Support. How is the altmetric score calculated? <https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-score-calculated-> (2015). [Online; accessed 12-Feb-2016]
33. H. Cunningham, Computers and the Humanities **36**(2), 223 (2002)
34. H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, in *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)* (2002)
35. J.R. Finkel, T. Grenager, C. Manning, in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (Association for Computational Linguistics, 2005), pp. 363–370
36. H. Cunningham, D. Maynard, V. Tablan, JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS-00-10, Department of Computer Science, University of Sheffield (2000). URL <http://www.dcs.shef.ac.uk/~diana/Papers/jape.ps>
37. W. Khawaja, M. Taylor, B. Davis, in *International Conference on Applications of Natural Language to Information Systems* (Springer International Publishing, 2015), pp. 443–447
38. A.E.M.C.P. Elkany, in *Proc of the SIGMOD*, vol. 1 (1997), vol. 1, pp. 997–23
39. S. Milojević, Journal of Informetrics **7**(4), 767 (2013)
40. A.M. Petersen, O. Penner, EPJ Data Science **3**(1), 1 (2014)
41. L. Page, S. Brin, R. Motwani, T. Winograd, (1999)
42. J.M. Kleinberg, Journal of the ACM (JACM) **46**(5), 604 (1999)
43. L. Katz, Psychometrika **18**(1), 39 (1953)
44. D.G. Bonett, T.A. Wright, Psychometrika **65**(1), 23 (2000)
45. J. Priem, H.A. Piwowar, B.M. Hemminger, arXiv preprint arXiv:1203.4745 (2012)
46. L. Waltman, R. Costas, arXiv preprint arXiv:1303.3875 (2013)
47. Z. Zahedi, R. Costas, P. Wouters, Scientometrics **101**(2), 1491 (2014)
48. M. Thelwall, S. Haustein, V. Larivière, C.R. Sugimoto, PloS one **8**(5), e64841 (2013)
49. J.H. Steiger, Psychological bulletin **87**(2), 245 (1980)
50. M. Kearns, D. Ron, Neural Computation **11**(6), 1427 (1999)
51. I. Jolliffe, *Principal component analysis* (Wiley Online Library, 2002)
52. L. Egghe, Journal of the American Society for Information Science and Technology **58**(3), 452 (2007)