Predicting Citations from Mainstream News, Weblogs and Discussion Forums

Mohan Timilsina
Insight Centre for Data Analytics
National University of Ireland Galway
mohan.timilsina@insight-centre.org

Brian Davis
Insight Centre for Data Analytics
National University of Ireland Galway
brian.davis@insight-centre.org

Mike Taylor
Digital Science
London, United Kingdom
mi.taylor@digital-science.com

Conor Hayes
Insight Centre for Data Analytics
National University of Ireland Galway
conor.hayes@insight-centre.org

ABSTRACT

The growth in the alternative digital publishing is widening the breadth of scholarly impact beyond the conventional bibliometric community. Thus, research is becoming more reachable both inside and outside of academic institutions and are found to be shared, downloaded and discussed in social media. In this study, we linked the scientific articles found in mainstream news, weblogs and Stack Overflow to the citation database of peer-reviewed literature called Scopus. We then explored how standard graph-based influence metrics can be used to measure the social impact of scientific articles. We also proposed the variant of Katz centrality metrics called *EgoMet* score to measure the local importance of scientific articles in its ego network. Later we evaluated these computed graph-based influence metrics by predicting absolute citations. Our results of the prediction model describe 34% variance to predict citations from blogs and mainstream news and 44% variance to predict citations from Stack Overflow.

CCS CONCEPTS

• **Information systems** → *Web mining*; *Social networks*;

KEYWORDS

Graphs, Centrality, Impact, Prediction, Altmetrics

ACM Reference format:
Mohan Timilsina, Brian Davis, Mike Taylor, and Conor Hayes. 2017. Predicting Citations from Mainstream News, Weblogs and Discussion Forums. In *Proceedings of WI '17, Leipzig, Germany, August 23-26, 2017,* 8 pages.
https://doi.org/10.1145/3106426.3106450

1 INTRODUCTION

The latency of traditional bibliometric indicators has led to the development of novel, alternative measures called Altmetrics [22]. Altmetrics can refer to (i) metrics for measuring alternative scientific artefacts e.g. source codes or datasets (ii) the measuring of impact beyond conventional boundaries of the scientific community. With the advent of the web and digital publishing and distribution, the audience for scientific work has broadened to include non-specialists. In the case of conventional citation-based metrics, if a paper has been cited then it can be judged to have had some scientific influence. However, it is less clear what influence can be established when a tweet about this paper is made from a non-specialist. Furthermore, not all scientific topics, however, scientifically excellent, may be accessible to a popular audience. So, it is hard to gauge the impact of scholarly activity in social media. According to a definition provided by Kaplan and Haenlein [13] social media means those Internet-based applications which allow creating, exchanging and disseminating the user generated content online on the ideological and technological foundations of Web 2.0. 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, and hyperlinks [21, 26]. The other important aspect of the Altmetrics is the choice of data sources. Many of the Altmetrics research is focused on analyzing the specific data sources particularly *Twitter*, as it reflects a wider use of scholarly articles by the general public [8]. Similarly, Twitter provides excellent API[1] to extract data for analysis. However, Twitter has also some limitations, the top tweeted scientific articles are with funny titles or curious stories [11]. These stories get higher attention and receive a higher number of tweets and retweets counts. These resonating count scores in media like Twitter and Facebook can be gamed or manipulated. To address this apparent weakness in such media, we chose three different data sources namely mainstream news, weblogs, and Stack Overflow[2]. These platforms provide lengthier and authoritative discussion about a particular topic. We support when a scientific publication is mentioned or linked in such media, they are more likely to be impactful in a social context.

Not all the scientific publication is featured in social media. The general presence of Altmetrics in "Biomedical and Health sciences" is 22% and for "Mathematics and Computer Science" is 5 % [8]. This infers that the public health and life sciences stories get much publicized in social media in comparison to other areas of sciences. To

WI '17, August 23-26, 2017, Leipzig, Germany

ACM ISBN 978-1-4503-4951-2/17/08…$15.00
https://doi.org/10.1145/3106426.3106450

[1] https://dev.twitter.com/rest/public
[2] http://stackoverflow.com/

measure the impact of scientific articles around public health stories we chose mainstream news and weblogs because newsworthy scientific topics might get serious attention from the public. For less visible stories we chose Stack Overflow mathematics community. Stack Overflow is not only focused Question&Answer community for a productive learning environment but also steady discussions about the topic. This community has the significant fraction of the participants having deep expertise in the domain area. Any scholarly articles linked in Stack Overflow gets score by the community members on the basis of how *useful* or *informative* they are [3].

The count based metric around scientific publication on social media are measuring the attention in online media, but it is hard to estimate that count metrics is really measuring the impact of science because any controversial or catchy title of scientific publications can get high counts. This metrics can be sensitive towards the popular trends called as popularity, but it does not measure the qualitative aspect for example prestige. The concept of popularity and prestige are established metrics in social network analysis [30].

Thus in this paper, we explore different graph-based influence metrics to assess the impact of the scholarly articles in an online social media. Then, we measure the impact of scientific publications in two different graphs: hyperlink document graph in mainstream news/weblogs and user activity graph in Stack Overflow. We evaluated the computed graph-based metrics for predicting the academic impact.

2 RELATED WORK

The traditional citation graphs proved notable successes to capture the important properties of the underlying research system. These graphs are appropriate to recognize the influence of bibliometric entities like scholars and journals [5, 29]. Similarly, the heterogeneous network studies from [33] and future rank algorithm proposed by [32] are applied to conventional citation and co-authorship network. However, Altmterics aimed at developing impact metrics of science in social media. The social media exhibit a rich variety of information sources, but also contain links between them [1]. This makes social media as a graph where documents are nodes and edges are hyperlinks. There are few studies around measuring the impact of science in social media using graph-based centrality approach [12, 17]. The limitation of [12] is the small sample size of only 45 researchers and the data are only from the academic social network called ResearchGate[3]. The findings of their studies can be biased because researchers can use different Online Social Networks such as Twitter, Facebook, Mendeley, Blogs, etc.

Hit Count [19] is a metric which captures the number of times a publication accessed online. It is used as a predictor to predict the citation count of medical scientific publications with variance of 33%. The study by Callaham [7] found that the "Newsworthiness" of the medical scientific article is the important predictor of the future citation count. However, Kulkarni et al [16] reported that Newsworthiness of medical literature has no significant association with citation rate. Brody et al. [6] presented web usage (number of downloads from the pre-print sharing website "arXiv") to predict the scientific impact (citations) of research articles. Similarly, Shuai et al. [24] investigated the relationship between Twitter mentions, arXiv downloads, and article citations using regression and correlation test and reported Twitter mentions is statistically correlated with arXiv downloads and early citations enable to use Twitter mentions and arXiv downloads to predict citations. Eysenbach [9] showed that the scientific articles mentioned on Twitter can predict future citations. The author reported, publications from Journal of Medical Internet Research (JMIR) which were highly tweeted were more likely to become highly cited. Ringelhan et al. [23] studied any unpublished scientific articles receiving likes in Facebook as an early indicator to predict the impact of scientific work.

Most of the prediction analysis have been performed on the bibliometric data sets [34] but few of the initiative were taken to predict the scientific impact using social media data. In this work, we focus on blogs, mainstream news, and Stack Overflow because these media bring attention to research output than any other social media [28]. To the best of our knowledge, there is no such graph-based approach in social web data to measure the social impact of scientific articles and predict the academic citations.

3 EXPERIMENT SETUP AND METHODOLOGY

We investigated three research questions. First, we examined whether we can identify the scientific sources in social web data. Second, we investigated the centrality metrics of identified scientific publications in such network. Finally, we evaluated the the computed graph-based centrality metrics by predicting academic citations.

RQ:1 How can we identify the scientific literature in a social web data?

We used two data sources in this study

(i) Mainstream News and Blogs Data: We used *Spinn3r*[4] data which is a crawl of the blogosphere from 2010 November to 2011 July. The data was stored in a distributed file system and has eight publisher types: *memetracker, forum, microblog, review, classified, mainstream news, weblog and social media such as facebook and twitter*. We extracted only weblogs and mainstream news from these distributed file using Java *Spinn3r* API[5] and stored in a MongoDb[6] database. We indexed extracted data using Solr[7] for quick search of the topic of interest.

Search of a Candidate Topic: 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 Wikipedia to research 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 [10] engine suggests bursts in Web user interest in these topics. We created a subset of the data for our focus topic from Spinn3r. For this, 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". We collected 259,149 JSON documents from Spinn3r dataset.

Construction of Spinn3r Graph: We constructed the hyperlink graph from the Spinn3r data by following the graph model [27]

[3] http://www.researchgate.net/

[4] http://spinn3r.com/

[5] http://www.programmableweb.com/api/spinn3r

[6] https://www.mongodb.org/

[7] http://lucene.apache.org/solr/

made for Targeted Project at Insight Centre for Data Analytics[8]. Each Spinn3r data item has the source URL and content. In the content section of the every item, we searched for the hyperlinks and from each hyperlink, we extracted the URL and these Urls are the target URL. We constructed a directed graph with source nodes as a source URL and target nodes as target URL. The graph contains 949611 number of nodes and 5408825 number of edges.

(ii) Stack Overflow Data[9]: We chose Math Overflow community from Stack Overflow because users contribute research level math questions and answers [25] which is ideal for our case. For our analysis, to be of reasonable size we restricted our data from January to December 2010. The retrieved size of the data was 345 MB.

Construction of User Activity Graph from Stack Overflow Data: From the Stack Overflow data, we made a graph with two relationships namely, "comment" and "share". We extracted those users who post the question and who response the post with comments. We linked these users with "comment" relationships. For the "share" relationship, we extract those users who shared the hyperlinks . We link the users and hyperlinks with "share" relationship. We stored this graph in a Neo4j [10] graph database. 713 user nodes, 518 hyperlink nodes, 1397 "comment" relationships and 515 "shares" relationships were created from this graph. Figure 1 shows the graph of the interaction between users sharing the scientific publications in Math Overflow community.

Identification of Scientific Source Domains in a Spinn3r and Stack Overflow graph: We took a semi-automated approach to identify the scientific publication link in the graph. We made the list of possible URL that can be found online from different academic search engines[11] because they cover the scientific disciplines in online social media [20]. The list contains Google Scholar, ScienceDirect, Nature, Science, New England Journal of Medicine (NEJM), The Lancet, PubMed, IEEE Xplore, arXiv, CiteSeer, Public Library of Science (PLOS) and Direct Digital Object Identifiers (DOI) (Table 1)

We used the list of the URLS shown in the table 1 and performed the following steps:

- **DOI and PMID Approach:** The scientific publications can be represented in the web as a unique persistent identifier called Digital Object Identifier(DOI)[12] such as http://dx.doi.org/10.1371/journal.pmed.1000388. Then DOI of this publication is 10.1371/journal.pmed.1000388. Similarly, PMID is the unique identifier used in PubMed Citations[13] for example https://www.ncbi.nlm.nih.gov/pubmed/3472723 then PMID is 3472723. We searched the URL's with the pattern **dx.doi.org** and **ncbi.nlm.nih.gov/pubmed/** in our graph and extract DOI and PMID. We checked each of the identified DOI's and PMID's in a Scopus[14] database using Scopus API[15]. This API directly provides the flexibility to search the scientific publications using DOI and PMID.
- **Title Based Approach:** For the rest of the URL's without DOI and PMID we searched the pattern of the URL from a list of academic resources shown in table 1. For every matched URL in the graph, we visited the web page and extract the title of the publications. We searched the exact title in Scopus database using Scopus API. The URL with identified scientific publication in Scopus was only analyzed.

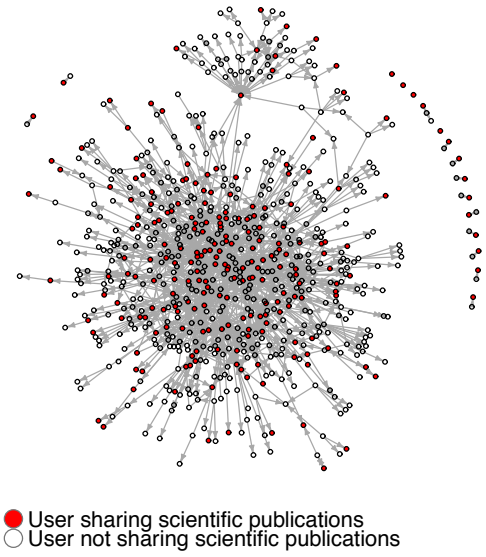

Figure 1: *User Interaction in Maths Community Sharing, Scientific Publications*

With this approach, we found 1210 scientific publications in a Spinn3r graph and 264 scientific publications in Stack Overflow graph. The sources of the scientific publications and their frequency are shown in Figure 2. For the Spinn3r graph, we observed the highest number of links (505) consist of a direct URL link to digital object identifier shown by (dx.doi.org). The second highest links are from library.wiley.com (420). The third highest links to scientific publications are linked through NCBI[16] pubmed (121). Similarly, in Math Overflow graph we observed the highest number of links (174) which were from arxiv.org. The second highest links (85) were from a direct URL link to the digital object identifier. The the third and the last were from library.wiley.com (3).

Finally, we constructed Spinn3r and Math Overflow graph and identified the scientific nodes in it. In the next section we address our second Research Question.

[8] https://www.insight-centre.org/
[9] http://stackoverflow.com/
[10] https://neo4j.com/
[11] http://www.sciencebuddies.org/science-fair-projects/top_science-fair_finding_scientific_papers.shtml
[12] http://www.apastyle.org/learn/faqs/what-is-doi.aspx
[13] http://answers.library.curtin.edu.au/faq/121100
[14] https://www.scopus.com/
[15] http://dev.elsevier.com/
[16] https://www.ncbi.nlm.nih.gov/

Sources	URL	Disciplines
Google Scholar	https://scholar.google.com/	All
ScienceDirect	http://www.sciencedirect.com/	All
Nature	http://www.nature.com/	All
Science	http://science.sciencemag.org/	All
Lancet	http://www.thelancet.com/	Medicine
NEJM	http://www.nejm.org/	Medicine
PubMed	https://www.ncbi.nlm.nih.gov/pubmed/	Life Sciences
arXiv	https://arxiv.org/abs/	Physics, Mathematics, Computer science, Quantitative biology, Quantitative finance and statistics
CiteSeer	http://citeseerx.ist.psu.edu/	Computer Science
Public Library of Science (PLOS)	https://www.plos.org/	Life Sciences
DOI system	http://www.doi.org/	A unique persistent identifier for online access of scientific publication
IEEE XPlore	http://ieeexplore.ieee.org/	Electronics, Electrical engineering, Computer science

Table 1: List of Academic Resources for Various Scientific Disciplines

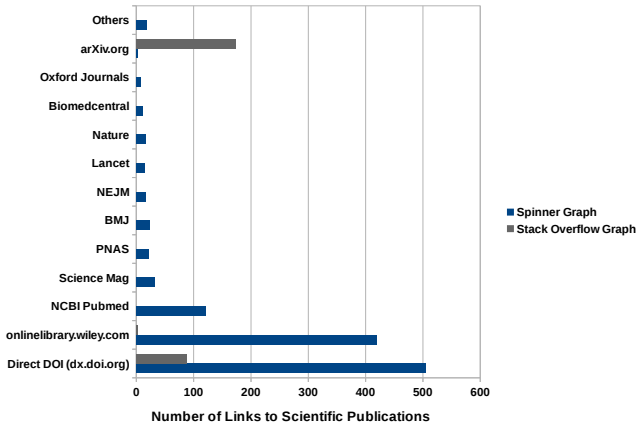

Figure 2: *Social Media Sources Citing Scientific Publication*

RQ:2 How can we measure the impact of scientific publication that are linked in Spinn3r and Math OverFlow graph? We applied four different centrality metrics out of which, three are standard graph-based centrality metrics namely PageRank [18], HyperLink Induced Topic Search (HITS) [15] and Katz Centrality [14]. The last one is called EgoMet, it is the metric we proposed. The algorithm such as PageRank, HITS, and Katz are established graph-based metrics which measures the global importance of the nodes in the network. However, these metrics may not be useful for determining the relative importance of the nodes with respect to a specifically focused root node [31]. The proposed metric is a variant of Katz centrality metrics to measure the influence of the root node in the focused ego-graph. In our analysis, we used all these metrics to assess the importance of scientific publications in both the graphs.

PageRank Score: For a given network of scientific publication and URL entries in a Spinn3r graph connected through the hyperlinks, the PageRank score of the scientific publication is the probability for a random surfer to land on it following these hyperlinks. We computed the PageRank score of 1210 scientific publications which are hyperlinked in the Spinn3r graph.

In the context of Math Overflow graphs, we applied the idea of combined PageRank score. For a given network of scientific publication and users connected through the comment and share relationship, the PageRank score of the scientific publication is the probability for a random surfer to land on it following comment relationship. In other words, if the users sharing the scientific publications are easily accessible then there is a high chance to reach that scientific publications. For any user nodes to accumulate higher PageRank score, it must be linked by other users with higher PageRank score (fig 3).

We summed up the PageRank score of the users who shared the scientific publications to provide one composite score for the publication. In order to compute the centrality of publication with n user based on PageRank score is given by:

$$Impact_{(Publication_{PR})} = \sum_{i=1}^{n} User_{(PR_i)} \quad (1)$$

where PR is the PageRank score of a user and n is the number of users who shared scientific publications.

Authority Score: We applied the HITS authority algorithms in our Spinn3r graph to compute the authority score of 1210 scientific publications. For a given network of scientific publication and URLs of blogs and mainstream news connected through the hyperlinks, the scientific publications have high authorities if they do not have the outgoing hyperlinks to the other URL entries.

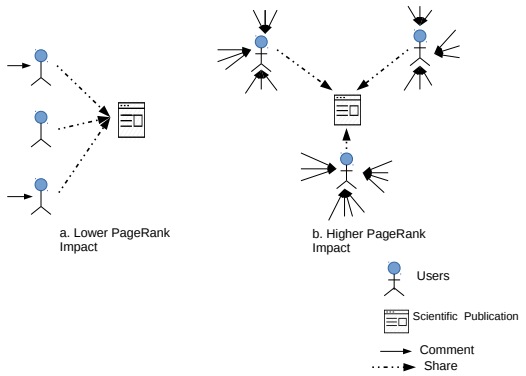

Figure 3: ***PageRank Score of Scientific Publications Shared by Users in StackOverflow Graph***

In the context of Math Overflow graphs, we applied the idea of combined Authority score similar to the concept described for combined PageRank score. To compute the centrality of publication with n user based on Authority score is given by:

$$Impact_{(Publication_A)} = \sum_{i=1}^{n} User_{(Authority_i)} \tag{2}$$

where A is the Authority score of a user and n is the number of users who shared scientific publications.

Katz Score: Katz centrality gives the relative influence of the nodes in the network. The measure of this centrality depends upon the number of immediate and distant neighbors. The nodes that lie very close to many other nodes in the network have a higher score than nodes lying farther from all the other nodes in the network. We applied this algorithms in our Spinn3r graph to compute the influence of 1210 scientific publications.

In the context of Math Overflow graphs, we applied the idea of combined Katz score similar to the concept described earlier for combined PageRank score and Authority score. To compute the centrality of publication with n user based on Katz score is given by:

$$Impact_{(Publication_K)} = \sum_{i=1}^{n} User_{(Katz_i)} \tag{3}$$

where K is the Katz score of a user and n is the number of users who shared scientific publications.

EgoMet Score: This is our proposed metric to weight the nodes in a maximal directed ego network. The root or ego node is the scientific publications and the other nodes are the set of alters who have ties to ego. The definition of Directed Ego-Centered Network and Maximal Directed Ego Network is given as:

Definition 1: *Directed Ego Centered Network*: For a graph $G = (V, E)$ where V is the set of nodes and $E \subseteq VXV$ 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.

The Katz centrality score has the attenuation parameter that describes how the signal decays when it traverses through hops. In the context of our graph we computed the attenuation parameter α [4] by

$$\alpha = \frac{1}{\lambda} \tag{4}$$

Where λ is the maximum eigenvalue of the graph. Similarly, the Node weight is computed as the ratio (R_{IO}) of indegree ($deg^-(Node)$) to outdegree ($deg^+(Node)$). R_{IO} provides the information spreading ability of the nodes. Mathematically, it can be represented as:

$$R_{IO}(Node) = log\left[\frac{deg^-(Node)_i+1}{deg^+(Node)_i+1} + 1\right] \tag{5}$$

The rationale to use log is for a very high indegree of the 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. We combine both the parameters to give one composite score called EgoMet score. Formally, it is expressed as follows:

$$EgoMet = \sum_{i=1}^{n} log\left[\frac{deg^-(Node)_i+1}{deg^+(Node)_i+1} + 1\right] * \alpha^{i-1} \tag{6}$$

Where n is the number of hops in a maximal directed ego network, i is the number of nodes in every hop and α is the attenuation parameter. This concept is demonstrated in figure 4 by computing the paper influence in 5 different network configurations.

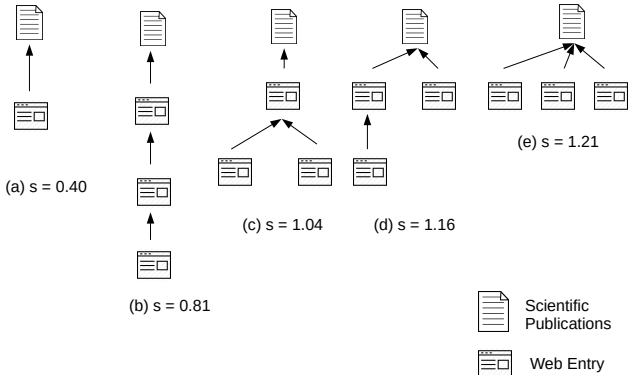

Figure 4: ***EgoMet Scores for 5 network configurations, with α=0.156. The score of the scientific publication grows from (a) to (e)***

In the context of Math Overflow graph, we used the combined EgoMet score of a user who shares the scientific publications. The combined centrality score of a scientific publication based

on EgoMet score of a user is computed as:

$$Impact_{(Publication_{EgoMet})} = \sum_{i=1}^{n} User_{(EgoMet_i)} \quad (7)$$

Finally, we applied 4 different graph-based metrics to assess the influence of scientific publications in Spinn3r and Stack Overflow graph. In the next step, we tried to answer our third research question.

RQ:3 How can we evaluate the computed graph-based metrics of scientific publication?

We performed the spearman correlations between the computed graph-based scores with the citation count baseline to find the relationship between these two scores. The baseline citation data were extracted from the time range of 2010 - 2011 because our Spinn3r and Stack Overflow data was collected in the same time period. The result of correlation test are shown in Table 2

Metrics	Correleations from Spinn3r Graph	Correlations from StackOverflow Graph
PageRank Score	0.25 ***	0.29 ***
Authority Score	0.19 ***	0.26 ***
Katz Score	0.45 ***	0.25 ***
Indegree Score	0.38 ***	0.28 ***
EgeMet Score	0.38 ***	0.34 ***

*** indicates highly sigificant and $p < 0.05$.

Table 2: Correlation Between Computed Graph-Based Scores with Citation Baseline

We observed a positive and significant correlation in our sample. This means the graph-based scores of scientific publication and its baseline citation score move in the same direction. This suggest a prediction model can be established between the predictor variable (graph-based scores) and response variable (baseline citations). We made two separate prediction model for Spinn3r and Math Overflow Data. The objective of this is to study which social media the baseline citation prediction is well performed. We included graph-based influence metrics along with other additional features to make better prediction.

Prediction Model for Spinn3r Graph: We analyzed the citation of scientific publication for the Spinn3r graph using multiple regression models. This model is based on citations of 1210 scientific publications. We transformed the dependent variable using natural log because our citation data was highly skewed. There were scientific publications in our datasets with zero citation, as the logarithm of 0 is undefined 1 is added to the citation count. Table 3 shows the significance of the predictors used in the model. The Authority Score and Katz score are not a significant predictor for citation count. The EgoMet (p-value =0.00486) score is moderately significant in predicting citations. The coefficient of the EgoMet score -0.019 represents a unit increase in an EgoMet score decreases the predicted citation by -2%. This is because our model is log transformed, a unit increased in an independent variable output coefficient times the 100% which is 1.9 approximately 2. This unit suggests that the EgoMet score negatively predicts citation scores. It can be the case like high public engagement can resonate the social

Variable	Coefficient	p-value
PageRank Score	0.016	0.00524 ***
Authority Score	0.005	0.4
Katz Score	0.22689	0.099
News_Referred	0.067	0.001175 ***
Blogs_Referred	0.0084	0.59
EgoMet Score	-0.019	0.00486 **
Depth of the Ego Network	-0.011	0.00364 ***
Number of Nodes in a k-hop of Ego Network	-0.24	0.759
Indegree of Scientific Pubs	**0.02**	0.000258 ***

*** indicates highly significant and $p < 0.05$.

Table 3: Model 1- Significance of Predictors for Spinn3r Graph.

influence score but does not actually contribute impacting the scientific score. Whereas scientific articles referred from mainstream news (p-value=0.001175) is highly significant to predict future citation. The unit increase in a News_Referred score increases the predicted citation by 6%. This infers that any newsworthy scientific topic from authoritative media source like mainstream news might be serious and a scientist might cite those scientific publications. The PageRank Score of scientific publications (p-value=0.00524) is a highly significant predictor for citation count. The unit increase PageRank score increases the predicted citation by 1%. The in-degree score of scientific publications (p-value=0.000258) is a highly significant predictor for citation count. A unit increase in the in-degree score increases the predicted citation by 2%. The high in-degree around the scientific articles might generate academic impact. The depth of the ego-network (p-value=0.00364) of the scientific publication also negatively predicts citations. This means the one unit increase in citation yield 1% decrease in a citation. The rest of the other predictors in the model are not statistically significant.

To check the prediction accuracy of the model, we split the datasets into training(75%) and testing(25%) set. This leads our training set sample to 908 and testing set the sample to 302. We choose RMSE (root mean squared error) as our evaluation metric because it gives high weights to the larger residuals. Residual is the difference between actual and predicted value. The baseline RMSE of our model is 16.54. This is calculated by taking the square root of the square difference between mean predicted value in training set and actual value in a test set. The R-squared (R^2) value of the model is 0.34. This means that 34% of the variance in our citation score can be explained by the set of predictors in the model. The RMSE value of the model while predicting in the test set is 10.40. The RMSE value in the test set is lesser than the baseline RMSE. This means our model predicts better than the baseline model.

Prediction Model for Math Overflow Graph: We followed the similar approach as a prediction model for a Spinn3r graph. We used the multiple regression models for Stack Overflow graph by using natural logarithm for the citation data. This model consists of a citation of 264 scientific publication. Table 4 shows the significance of the predictors used in the model. The EgoMet score (p-value = 1.92e-07), Authority Score (p-value = 1.23e-06) and PageRank Score (p-value = 0.000836) are three graph-based metrics which

Variable	Coeffecient	p-value
PageRank Score	0.0299510	0.00836 ***
Authority Score	0.098590	1.23e-06 ***
Katz Score	-0.08996	2.38e-05 ***
EgoMet Score	0.27011	1.92e-07 ***
Maximum Hops of the Ego Network	-0.002344	0.889
Indegree of Scientific Pubs	0.05876	0.00176 **
Number of Answers	0.13749	8.91e-07 ***
Number of Votes Sharing Scientific Publications	0.01130	0.000052***
Number of Comments	-0.001	0.92

*** indicates highly significant and p < 0.05.

Table 4: Model 2 - Significance of Predictors for Math Overflow Graph.

Sources	Model	Data	R^2
Perneger (2004)	Linear Regression	Online web access data of the scientific publications.	33%
Brody (2006)	Linear Regression	Web usage data of scientific publications	42%
Eysenbach (2011)	Linear Regression	Twitter	27%
Yan, Rui, et al (2011)	Classification and Regression Tree	Bibliometrics data	78%
Our Model 1	Linear Regression	Mainstream News and Blogs	**34%**
Our Model 2	Linear Regression	Math Overlow	**44%**

Table 5: List of Different Predictive Models to Predict Academic Citations

are highly significant predictors to predict the citations. The one unit increase in Egomet score, Authority score, PageRank score increases citations by 27%, 9% and 2 % respectively. Likewise, Indegree (p-value = 0.00176) of scientific publications is also a significant predictor of citations. A unit change in an Indegree increases citations by 5% . The number of votes sharing scientific publication is also highly significant feature in the model. This feature suggests a unit change in votes increases citations by 1 %. The number of the votes for scientific publication shared in the Math Overflow might infer usefulness of the resource. Thus it can be taken as an important indicator to predict citations. Similarly, the number of answers (p-value = 8.91e-07) posted indicates highly significant predictor in the model. For a unit change in a number of answers increases citations by 13% . The number of comments PageRank score and Maximum hops of the ego network is not statistically significant feature for the prediction of citations in the model.

To check the prediction accuracy of the model, we split the data into training(75%) and testing(25%) set. Our training sets consist of 198 entries and test sets consist of 66 entries. The baseline RMSE of this model is 7. The R-squared value of the model is 0.44. This means our model can explain 44% variance in the citation. The RMSE value of the model for test set is 5.07. This value is lesser than our baseline RMSE value of the model which suggest the model predicts better than the baseline model.

Comparison of different Models Predicting Citation Counts: We took four different models where three of the model predict the citations using social web data and one model which predict citations using bibliometric data. The performance based on the coefficient of determination (R^2) is shown in Table 5. In comparison with all the model, we observed that the R^2 value is maximum for the model from Yan Rui,et al (2011). This model explains the maximum variance of the academic citations. The reason for this is the author used all the predictor variables which are bibliometrics indicator such as author ranks and venues rank which are highly correlated with the academic citations. Considering all the features used by social media, it shows that our model 1 offers slightly better variance than Perneger (2004) and Eysenbach (2011). Similarly, our model 2 describes the highest variance in comparison to other models. The model from Brody (2006), Perneger (2004) and Eysenbach (2011) are based on counts. For example, counts of downloads, mentions, and exposure of scientific publications in an online medium. These count based metrics are important because it captures the popularity or attention but it does not capture influence. Whereas our graph-based metrics capture the influence of scientific publication in social media.

The prediction model built from Math Overflow describe higher variance in citation than prediction model build from Spinn3r data. One reason for this might be the users. In the specialized and dedicated forum of mathematics in Math Overflow the users might be matured or novice scientist or normal audience. Any scientific publication shared in such forum might get attention and if useful it might also get cited. The other important aspect of Math Overflow is that it has reward system on the basis of trustworthiness and accuracy of the content [2]. This means any scientific publication shared by users get votes. We also saw from our prediction model that the number of votes is significant predictors in predicting citations. In the case of mainstream news and weblogs in Spinn3r data, the majority of the audience might be non-specialist. For such users, they might read the news or blogs linking scientific publication which contains catchy headlines or interesting stories. These users might have less chance to visit the primary source or scientific publications. This factor actually resonates the social scores but does not contribute much to academic citations.

Although we were not very sure about how our approach performs to predict academic citations with other similar studies. The reason for this is we do not have the experimental data used by those studies to reproduce the experiment. But we present the coefficient of determination (R^2) from those studies reported in the literature. We compared this with the previous studies because they used social media to predict academic citations.

4 LIMITATIONS AND CONCLUSION

The limitation of the study is that the number of scientific articles in Spinn3r (mainstream news and weblogs) and StackOverflow are 1210 and 264 respectively which is of small size and incomplete. This is due to the method we used to curate scientific articles from social media. We have done this manually by identifying URL's from a scientific domain and matching them in Scopus database. While doing this, we might have missed other potential scientific resources in social media. Due to this, the approach of recognizing scientific articles in social media needs to be automated. As in our use case, we chose widely-publicized public health issues called *"avian flu"*. This kind of sensitive stories might bias the finding or

prediction. This needs to be further verified for all kind of scientific publications available in social media. This can be a potential future direction for this research.

Our work is an exploration for looking at the academic impact of scientific publications outside the conventional bibliometric community. We linked the scientific publications found in social media to peer-reviewed literature database and measured the social impact of such publications. We also proposed the EgoMet score to measure the local influence of the nodes in the ego network which showed a moderate correlation and positive association with a citation baseline. Finally, we evaluated the computed graph-based metrics by predicting academic citations.

ACKNOWLEDGMENTS

We would like to acknowledge Science Foundation of Ireland (SFI/12/RC/2289) and the targeted project Elsevier for funding this research.

REFERENCES

[1] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining*. ACM, 183–194.
[2] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2012. Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 850–858.
[3] Nina Belojevic, Jentery Sayers, et al. 2014. Peer review personas. *Journal of Electronic Publishing* 17, 3 (2014).
[4] Phillip Bonacich and Paulette Lloyd. 2001. Eigenvector-like measures of centrality for asymmetric relations. *Social networks* 23, 3 (2001), 191–201.
[5] Sergey Brin and Lawrence Page. 2012. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks* 56, 18 (2012), 3825–3833.
[6] Tim Brody, Stevan Harnad, and Leslie Carr. 2006. Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology* 57, 8 (2006), 1060–1072.
[7] Michael Callaham, Robert L Wears, and Ellen Weber. 2002. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *Jama* 287, 21 (2002), 2847–2850.
[8] Rodrigo Costas, Zohreh Zahedi, and Paul Wouters. 2015. 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, 10 (2015), 2003–2019.
[9] Gunther Eysenbach. 2011. Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of medical Internet research* 13, 4 (2011), e123.
[10] google keyword search 2016. Google Trend. https://www.google.com/trends/explore?date=2010-11-01%202011-07-31&q=%2Fm%2F0292d3. (2016). [Online; accessed 7-August-2016].
[11] Stefanie Haustein, Isabella Peters, Cassidy R Sugimoto, Mike Thelwall, and Vincent Larivière. 2014. Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *Journal of the Association for Information Science and Technology* 65, 4 (2014), 656–669.
[12] Christian Pieter Hoffmann, Christoph Lutz, and Miriam Meckel. 2014. Impact factor 2.0: Applying social network analysis to scientific impact assessment. In *2014 47th Hawaii International Conference on System Sciences*. IEEE, 1576–1585.
[13] Andreas M Kaplan and Michael Haenlein. 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons* 53, 1 (2010), 59–68.
[14] Leo Katz. 1953. A new status index derived from sociometric analysis. *Psychometrika* 18, 1 (1953), 39–43.
[15] Jon M Kleinberg. 1999. Hubs, authorities, and communities. *ACM computing surveys (CSUR)* 31, 4es (1999), 5.
[16] Abhaya V Kulkarni, Jason W Busse, and Iffat Shams. 2007. Characteristics associated with citation rate of the medical literature. *PloS one* 2, 5 (2007), e403.
[17] Na Li and Denis Gillet. 2013. Identifying influential scholars in academic social media platforms. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 608–614.
[18] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: bringing order to the web. (1999).
[19] Thomas V Perneger. 2004. Relation between online "hit counts" and subsequent citations: prospective study of research papers in the BMJ. *BMJ* 329, 7465 (2004), 546–547.
[20] Jason Priem, Paul Groth, and Dario Taraborelli. 2012. The altmetrics collection. *PloS one* 7, 11 (2012), e48753.
[21] Jason Priem and Bradely H Hemminger. 2010. Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday* 15, 7 (2010).
[22] Jason Priem, Heather A Piwowar, and Bradley M Hemminger. 2012. Altmetrics in the wild: Using social media to explore scholarly impact. *arXiv preprint arXiv:1203.4745* (2012).
[23] Stefanie Ringelhan, Jutta Wollersheim, and Isabell M Welpe. 2015. I like, I cite? Do Facebook likes predict the impact of scientific work? *PloS one* 10, 8 (2015), e0134389.
[24] Xin Shuai, Alberto Pepe, and Johan Bollen. 2012. How the scientific community reacts to newly submitted preprints: Article downloads, twitter mentions, and citations. *PloS one* 7, 11 (2012), e47523.
[25] Yla R Tausczik and James W Pennebaker. 2011. Predicting the perceived quality of online mathematics contributions from users' reputations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1885–1888.
[26] Mike Thelwall. 2012. Journal impact evaluation: a webometric perspective. *Scientometrics* 92, 2 (2012), 429–441.
[27] Mohan Timilsina, Brian Davis, Mike Taylor, and Conor Hayes. 2016. Towards predicting academic impact from mainstream news and weblogs: A heterogeneous graph based approach. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*. IEEE, 1388–1389.
[28] N Seth Trueger, Brent Thoma, Cindy H Hsu, Daniel Sullivan, Lindsay Peters, and Michelle Lin. 2015. The Altmetric score: a new measure for article-level dissemination and impact. *Annals of emergency medicine* (2015).
[29] Yujing Wang, Yunhai Tong, and Ming Zeng. 2013. Ranking Scientific Articles by Exploiting Citations, Authors, Journals, and Time Information.. In *AAAI*.
[30] Stanley Wasserman and Katherine Faust. 1994. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press.
[31] Scott White and Padhraic Smyth. 2003. Algorithms for estimating relative importance in networks. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 266–275.
[32] Xin-min Xiang. 2009. Futurerank: Ranking scientific articles by predicting their future pagerank. (2009).
[33] Ding Zhou, Sergey A Orshanskiy, Hongyuan Zha, and C Lee Giles. 2007. Co-ranking authors and documents in a heterogeneous network. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. IEEE, 739–744.
[34] Xiaodan Zhu, Peter Turney, Daniel Lemire, and André Vellino. 2015. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology* 66, 2 (2015), 408–427.