Mining Intellectual Influence Associations

Tejas Shah, Vikram Pudi

International Institute of Information Technology - Hyderabad, India
tejas.shah@research.iiit.ac.in, vikram@iiit.ac.in

Abstract. Within the social system of science, citation practices characterize social functions like the conferral of recognition upon the work of others as well as the acknowledgement of one's intellectual debt. However, the structure of intellectual influence is misrepresented when only the immediate citations and their cardinality are taken into consideration. Thus, in order to better understand the associative dissemination of influence and approximately construe the anatomy of this structure, complex interactions in the convoluted network of authors and papers need to be probed. Our study aims at understanding these heterogeneous complex interactions. For the bibliographic dataset of authors and publications, we define proxy scores that attempt to determine the associative influence of the cited author over the citing author. In order to harness structural connectivity of the network, we generate author vector representations using these influence scores. Furthermore, with a view to assess the competence of our proposed scores, we evaluate these representations and provide an empirical study of the results obtained with our algorithm against the baseline and also present a qualitative analysis.

Keywords: Citation Network, Influence, Representation Learning

1 INTRODUCTION AND RELATED WORK

The contribution of an author in the form of publications holds an intrinsic value responsible for the effective dissemination of knowledge. This knowledge based relay that is extended by and for the scientific community results into establishment of conceptual relationships in the form of citations. Citations and references operate within a jointly cognitive and moral framework designed to provide the historical lineage of knowledge and to repay the intellectual debts through their open acknowledgement [15][7]. Thus, citation analysis has the potential in providing valuable insights about the social system of science spanning across a wide range of topics.

Effective digitization of bibliographic data has led to proliferation in the propositions of various quantitative bibliometric performance indicators for journals, papers, authors and institutions based on citation counts and graph-based ranking algorithms [6][17][22]. Studies dealing with influence have examined various aspects like topic level influence strength, influence propagation and its indirect global effect [8]; analysis of influence evolution between communities [4];

time dependent estimation of influence for evaluating pairwise community influences [18]. Besides being extensively explored in literature, the concept of influence has also surfaced in scholarly search engines such as Semantic Scholar. Many of these approaches focus on influence analysis based on individual entities and their overall impact on the network structures. However, influence relationships and associations do not emerge inherently considering just the global influence. These influence associations can be understood as the degree of influence between a pair of nodes within the network. Our work aims at studying these pairwise influence associations eventuated between authors within the scholarly network. The main contributions of the paper are as follows:

- We propose an algorithm that simulates the influence between the citing and cited authors and suggest influence association scores.
- Considering issues in quantifying and tracing of influences that arise in scholarly communication and to harness structural connectivity of the network, we profile authors and their interactions within the bibliographic network using representation learning and the proposed influence scores. These representations thus form a generic result of the proposed influence model and their effectiveness in context of the problem statement is discussed.
- For assessment of predictive capacity of the aforementioned scores and the thus obtained vector representations, experimental results subject to classification tasks are discussed along with comparative study against those obtained by measures such as citation counts.

2 PROBLEM STATEMENT

2.1 Problem Formulation

Consider a publication p_i written by m co-authors a_{i_1}, a_{i_2}, .., a_{i_m}, which cites a publication p_j written by n co-authors a_{j_1}, a_{j_2}, .., a_{j_n}. Thus, a publication citation network can be defined as a directed graph $G_P = (V_P, E_P)$ constructed from the list of references at the end of each publication, where V_P represents vertices (publications) and E_P represents set of all directed edges between the nodes denoting citations between publications. Consequently, an author citation network $G_A = (V_A, E_A)$ is defined by projecting this publication citation network along the corresponding author(s) for each publication node in G_P. For instance, the citation link $p_i \rightarrow p_j$ is projected between their authors respectively, thus creating $m \cdot n$ directed links from each of the m co-authors of the citing publication to each of the n co-authors of the cited publication. Accordingly, V_A represents nodes (authors) and E_A represents set of all such directed links between the authors. Let the directed pairwise author citation link between the citing author a_i and the cited author a_j be denoted as $a_i \rightarrow a_j (\forall i = 1, .., m$ and $\forall j = 1, .., n)$. In the discussions that follow, we define and aim to quantify the associative intellectual influence measure represented by $I(a_i, a_j)$ as the degree to which author a_j influences author a_i when a citation is made from a_i to a_j.

2.2 Issues and Caveats in Quantifying Influence

Citation networks are complex networks in which causal structure exists along the interactions between the nodes. However, consideration of just the citation counts and primary degree of interactions (immediate neighbours) within the citation network has its shortcomings as indicators for tracing intellectual influences [23]. Without a subjective survey of authors in conjunction with their publications, it remains unknown as to what fraction of the work is cited that was indeed influential directly or indirectly and whether the references exist which had no influence of any kind yet cited due to other motivations. Such complex interplay of multiple citer motivations have been empirically studied and reported in previous studies as well [2][10].

Possible classes of errors in tracing scholarly influences include:

- The inclusion (or exclusion) of a reference in a bibliography does not completely indicate whether or not those references were directly or indirectly influential for the proposition of the publication.
- Citing bias in favor of elite scientists or highly cited papers i.e. over-citation described as the "Matthew effect" [11].
- Under-citation of fundamental scientific work is possibly noticed due to the obliteration (of sources) by incorporation (OBI) in the established knowledge [10].
- When a relevant piece of scientific work is known through an intermediate publication, the intermediate publication serves as an intermediate influence. However, it may remain uncited.
- Nature of citation types as they can be further categorized into organic or perfunctory, evolutionary or juxtapositional, confirmative or negational etc [14].

These certain and other such classes of errors exist within the citation data due to under-inclusion and over-inclusion of references [2][10][23]. This makes the task of tracing influences more complex.

3 ASSOCIATIVE INFLUENCE MEASURES

We propose two associative influence scores for effectively capturing intellectual influence of the cited author over the citing author. The underlying principle encapsulates the reasoning that the citations towards an *influential* publication of the cited author and those from the *finest* works of the citing author are indeed significant. Further, the net influential impact over a publication can be fairly attributed as an aggregation of influences by all the cited publications. Considering citing author's temporal scholarly activity, associative influences are instantiated for each such citing publication wherein references are made to other publications. Thus, for a particular author pair (a_i, a_j), we consider each such instance wherein a_i cites a_j i.e. all such publications authored by a_i wherein a citation has been made to a publication authored by a_j. This instantiated

influence forms a component for the integral associative influence between the author pair. Since the associative influence is a directed mapping, the influence scores resemble the same notion of directedness. Thus, for a citation relationship $a_i \rightarrow a_j$, the proposed influence association score $I(a_i, a_j)$ represents the degree of influence cited author a_j has over the citing author a_i.

3.1 Ranking Publications

In order to capture the collective nature of scholarly influence, publication-level ranking is adopted (as opposed to researcher-level ranking). This mimics the spread of intellectual influence among researchers via their publications. Quite a few studies [6][21] investigate extensively this key issue of scientific credit diffusion by dissecting the credit diffusion mechanism underlying both researcher level and paper level graph-ranking methods. Their findings emphasize that scientific credit is fundamentally derived from citation information between papers rather than the derived researcher network. Our model for the influence dissemination within the heterogeneous network of authors and papers thus avoids the inaccurate allocation of scientific credit among researchers that potentially arises in graph-ranking methods.

PageRank [1] takes into account the number and quality of links while measuring the importance of entities within a network. Using PageRank over G_P, the importance of publications is thus measured considering the number of citations and reputation of the papers [9]. For the publication level ranking, we have:

$$PR(p_i) = \frac{(1-\alpha)}{N} + \alpha \cdot \sum_{j \rightarrow i} \frac{PR(p_j)}{T_{out}(p_j)} \tag{1}$$

where $j \rightarrow i$ implies paper p_j referring paper p_i, $PR(p_j)$ denotes the PageRank of paper p_j and $T_{out}(p_j)$ denotes the number of outbound links from paper p_j. With certain empirical studies, Chen et al. [3] showed that, scientific papers usually follow a shorter path of about average two links. This is in opposition to six hyperlinks for the web considering the individual surfer illustration as mentioned in the original study [1]. Accordingly, we set the damping factor α to 0.5 for the purpose of our studies as well.

3.2 Source based Influence - $I_S(a_i, a_j)$

The basis for this score is that, citations from the *significant* works of the citing author (a_i) are indeed relevant, considering the intellectual and cognitive influences. The citing author's significant works can be regarded as those scholarly works which have a relatively high PageRank among other works of the same author. The instantiated associative influence components resulting from citing author's citation activity sum up to form the net score. Thus, we calculate $I_S(a_i, a_j)$ as follows:

$$I_S(a_i, a_j) = \frac{\sum_k \frac{PR(p_{ik})}{T_{out}(p_{ik})}}{\overline{PR(p_{a_i})}} \qquad \forall p_{ik} \in p_{ik} \rightarrow p_{a_j} \tag{2}$$

where, p_{ik} denotes the k^{th} publication of the citing author a_i wherein a citation has been made to a publication authored by the cited author a_j, $PR(p_{ik})$ represents the PageRank of the citing publication p_{ik}, $T_{out}(p_{ik})$ denotes the number of outbound links (references) from publication p_{ik} and $\overline{PR(p_{a_i})}$ is the normalization factor which is the average PageRank value of all the publications authored by a_i. The normalization factor accounts for the variance in the ranks of citing publications. So, higher the normalized weight of each such component (i.e. higher the normalized PageRank of citing publication) and higher the cardinality of such components exchanged between a_i and a_j, higher is the associative influence.

3.3 Destination based Influence - $I_D(a_i, a_j)$

The underlying notion for devising this score is to calculate a measure of the extent to which the citing author tends to cite a set of authors weighted by their influential scholarly contributions in his/her publications. Extending the influence instantiation notion for I_D, the association components of influence comprise of the PageRank of the cited publication and the inbound links of that cited publication (distributing impact of a publication among the inbound citation references). Thus, we calculate $I_D(a_i, a_j)$ as follows:

$$I_D(a_i, a_j) = \frac{\sum_k \frac{PR(p_{jk})}{T_{in}(p_{jk})}}{\overline{PR(p_{a_j})}} \qquad \forall p_{jk} \in p_{a_i} \rightarrow p_{jk} \tag{3}$$

Here, p_{jk} denotes the k^{th} publication of the cited author a_j wherein a citation has been made from a publication authored by the citing author a_i, $PR(p_{jk})$ represents the PageRank of the cited publication p_{jk}, $T_{in}(p_{jk})$ denotes the number of inbound links to publication p_{jk} and $\overline{PR(p_{a_j})}$ is the normalization factor which is the average PageRank value of the publications authored by a_j. The disparity in the ranks of cited publications is accounted for by this normalization factor. So, more the normalized weight of each such association components and higher the cardinality of such components exchanged between a_i and a_j, higher is the associative influence.

3.4 Qualitative Analysis

Author and publication ranks of prominent scientists and young researchers span across a wide spectrum. For example, a highly cited publication of a prominent

scientist might be influential but in a modest way to each individual researcher in the scientific community at large. Also, the finest works of a young researcher possibly might have less inbound citations comparatively, however, the substantial influence of cited publications within such publications should not possibly be overlooked. Considering the different degrees of author and publication ranks, it can be seen from Equation (2) that significant works of young as well as even mediocre researchers are instrumental in contributing towards highlighting the cited author's influence. Thus, irrespective of the cited author's prominence, his/her influence over the citing author is conspicuous and visibly pronounced. Also, from Equation (3) it is evident that even scientists not belonging to the higher order of ranks receive their due scholarly attribution pertinent to the their respective notable scientific studies. Such notions of relative author impact and allocation of due credit persist across the spectrum for most of the researchers belonging to different degrees of author ranks. Normalization of associative influence measures as illustrated in sections 3.2 and 3.3 account for such variances along with taking into consideration the effect of accumulated advantage (i.e. Matthew effect [11]).

Associative pairwise influences eventuated due to scholarly contributions over time semantically imply that there is certain form of influence of the cited author over the citing author irrespective of the citation types. In this paper, we focus on the existence and extent of conceptual relationships formed between authors. However, the precise nature of influence maybe hard to quantify without the factual ontological citation representations. For alleviating issues concerning under-inclusion of references (as discussed in Section 2.2), capturing the network neighborhood and harnessing the structural connectivity, we profile authors and their interactions using representation learning and the proposed influence scores. This effectively maps the *latent* features within the citation network into a vector space.

4 AUTHOR REPRESENTATION LEARNING

Recent work in language modeling and representation learning such as Word2Vec [12] focuses on application of probabilistic neural networks which map words into vector spaces. The author vector representations are learned with a similar intuition as discussed in the following sections:

4.1 Random Walk on Weighted Directed Network

Using proposed influence measures, we model *weighted random walks* over the author citation network. These walks can be approximated as sentences in the context of language modeling. Analogous to recent researches [16], it represents a network as a "document". The motivation behind converting a graph into a series of text documents is: Word frequency in a document corpus and the visited node frequency during a random walk for a connected graph, both follow the power law distribution [16].

We sample random walks over the weighted directed author citation network. Considering the author citation network (G_A) with n author nodes (a_1, a_2, ..., a_n); w_{ij} represents the weight of the edge connecting nodes a_i and a_j where the edge weight is the influence score $I(a_i, a_j)$ (as derived from Equation 2 or Equation 3) for the author pair and therefore we have, $w_{ij} = I(a_i, a_j)$. Since the influence scores are non-negative, we have $w_{ij} \geq 0$. For the study of tracing influence associations, we prune self-citation loops and thus $w_{ii} = 0$. Now, for a given source node a_i, the transition probability that the author node a_j is chosen from the direct successors of a_i is proportional to the influence measure $I(a_i, a_j)$. This is computed as:

$$p_{a_i,a_j} = \frac{I(a_i, a_j)}{\sum_k I(a_i, a_k)} \tag{4}$$

where p_{a_i,a_j} represents the transition probability. For each source author node a_i, we simulate a weighted directed random walk $\mathcal{W}_{a_i}$. This sampling is a stochastic process consisting of author nodes $w^1_{a_i}$, $w^2_{a_i}$, ..., $w^n_{a_i}$ as random variables such that $w^j_{a_i}$ is a vertex chosen with transitive probability p_{a_i,a_j} from the direct successors of a_i. In our experiments we set the length of these walks to be fixed. For each source vertex, the random walk generator samples author nodes based on respective transition probabilities until a maximum length l (= 40) is reached. For the purpose of our study, we generate such weighted random walks γ (= 15) times for each author.

4.2 Representation Learning Framework

Modelling social structures and relationships within networks can be aligned with the optimization techniques used to model natural languages [16][5]. With the ordered sequence of nodes constructed using weighted random walks, we learn representations using Skip-gram model. This represents authors and the citation relationship shared between a pair of authors in an unsupervised manner. Based on distributional hypothesis, these representations are latent features that capture *neighbourhood* as well as *structural* influences in the citation network in a continuous low dimensional vector space. In effect, these representations encapsulate more information and relationships between authors than using just the immediate citations.

Skip-gram is a language model that maximizes the conditional co-occurence probability of words occurring within a predefined window [12]. Thereby, we have $f : V_A \rightarrow \mathbb{R}^d$ as the mapping function from nodes to feature representations that we aim to learn. Here d specifies the number of dimensions of the feature representation and f represents a matrix of size $|V_A| \times d$ parameters. Now, we try to optimize the likelihood function as formulated in Equation 5:

$$\max_f \sum_{j=i-w, j\neq i}^{j=i+w} \log \Pr(a_j \mid f(a_i)) \tag{5}$$

where w is the size of the window, $a_i \in V_A$ and $\Pr(a_j \mid f(a_i)$ is defined by the softmax function:

$$\Pr(a_j \mid f(a_i)) = \frac{exp(f(a_j) \cdot f(a_i))}{\sum_{k \in V} exp(f(a_k) \cdot f(a_i))} \tag{6}$$

Skip-gram assumes that inside the context, all nodes are independent of each other and are equally important. However, as seen from Equations 5 and 6, update step per node is proportional to $|V_A|$. This is computationally expensive for large networks (such as the author citation network in consideration). We approximate the optimization function using negative sampling [13].

Using the obtained resultant author vector representations, we validate the effectiveness of our proposed scores as discussed in the following sections.

5 PERFORMANCE MODEL

Limitations in the assessment of intellectual and cognitive influences prevail due to its subjective nature. However, despite the shortcomings of citation data, studies [23] assert that citations can be used as approximate proxy indicators of influence for the aggregates of authors and papers. Collectively, with measurable factors and practical limitations of the study, it can be fairly argued that if a publication proves to be relatively influential in the scientific work of an author, then he/she is quite definitive to have a higher relative citation ratio towards the cited influential authors. So, we utilize and bucket these relative citation ratios for author pairs as labels for classifying the extent of influence. Representations obtained using influence scores can be evaluated against baseline by a comparative study of citation prediction and its extent between author pairs. In our study, the purpose is to evaluate whether our proposed influence measures capture meaningful relationships. To do so, it suffices to test their relative capacity in citation prediction, and the absolute predictive accuracy is not the criterion being assessed.

5.1 Validating Influence Associations

To capture the semantic relatedness between the citing and cited author's influential relationship and in order predict weighted citation link between a pair of authors a_1 and a_2, we generate edge representation $e(a_1, a_2)$. This is done by defining a binary operator over the corresponding author feature vectors $f(a_1)$ and $f(a_2)$. Similar strategies have been successfully used in earlier studies for link prediction tasks [5]. We define i^{th} component of the edge representation $e_i(a_1, a_2)$ as concatenation of author feature vector components denoted by $f_i(a_1) \sqcup f_i(a_2)$. $e(a_1, a_2)$ spans across $\mathbb{R}^{2 \times D}$ as the author feature vectors $\in \mathbb{R}^D$ are concatenated.

These edge representations are now further used in training and evaluation for predicting the influence between a pair of authors. However, predicting the

presence of link between the author pair in the testset is necessary but insufficient to assess predictive capacity of the degree of influence. For evaluating the extent of influence more rigorously, we extend this binary classification link prediction evaluation to multi-label classification. Here, the edge representations are classified against labels viz., *Nil Influence (NI)*, *Slightly Influenced (SI)* and *Highly Influenced (HI)* depending upon the relative citation ratio between the pairs of authors in the training and testing sets respectively. Thus, for a pair of authors a_i and a_j, $e(a_i, a_j)$ is classified based on the relative citation ratio ($cr_{(a_i,a_j)}$) which is computed as:

$$cr_{(a_i,a_j)} = \frac{\left|c(a_i, a_j)\right|}{\sum_k \left|c(a_i, a_k)\right|} \tag{7}$$

where $\left|c(a_i, a_j)\right|$ represents number of citations from the citing author a_i to the cited author a_j in the author citation network during that specific temporal segment. The calculated citation ratios are then mapped to aforementioned class labels. If $cr_{(a_i,a_j)} \in (0.0, \delta]$, then $e(a_i, a_j)$ is classified as SI. When $cr_{(a_i,a_j)} > \delta$, then $e(a_i, a_j)$ is classified as HI. Based on repeated experiments to maximize discrimination among citation ratios, δ is set to 0.036. Lastly, $cr_{(a_i,a_j)} = 0$ implies that there is no influence of the cited author a_j over the citing author a_i, thus, classifying $e(a_i, a_j)$ as NI.

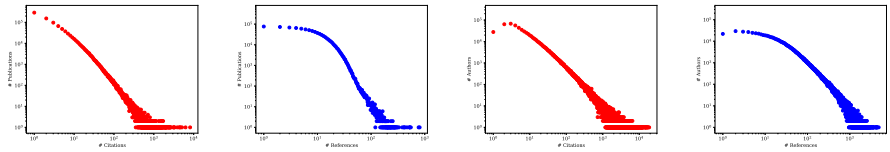
Baseline: For comparing the performance of our model and the influence scores, we use citation counts as baseline for our evaluation. Author profiling and generation of vectors for this baseline is achieved using weighted random walk over author citation graph. Here, weights are the number of citations from the citing author to the cited author (as opposed to influence measures as discussed in 4.1). Edge representations for this baseline are then computed using the obtained author vector representations.

6 EXPERIMENTAL DESIGN

6.1 Dataset Description

The DBLP dataset used consists of papers published from the period 1960 to 2014 wherein the citation data is enriched by using bibliographic metadata from ArnetMiner [20]. The dataset contains information such as paper's title, its authors and their affiliations, citation list, publication year, etc. For our experiments, pre-processing is performed over the dataset for pruning incomplete records where: 1) The publications with incomplete meta data (absence of year, authors, etc.) are removed. 2) Internal citations can be defined as references to publications within the snapshot of dataset being considered. References other than these internal citations are removed.

The final dataset includes 1277594 papers and 1003387 authors. The total count of internal references for publication citation network is 7962820 whereas edgelist for the author citation network sums up to 39713499.



(a) (b) (c) (d)

Fig. 1: Figures (1a), (1b) represent publication in-degree and out-degree distributions for the publication citation network and figures (1c), (1d) denote the author in-degree and out-degree distributions for the author citation network respectively.

6.2 Dataset Segmentation

In order to evaluate and assess the degree of influence as quantified by the suggested influence measures, we consider work of individual authors over a temporal scale. We divide the dataset into 3 segments as follows:

`Profiling`: This segment of dataset represents the activity of researchers and scientists within the bibliographic network up to 2006. Interactions between researchers by means of collaborations, citations and conceptual exchanges in the form of publications are significantly eventuated considering such a wide temporal span. This partition of dataset helps in the process of author profiling by means of learning author vector representations using weighted random walks.

`Training`: This segment of dataset is used for learning influence associations by generating edge representations using the author vector representations captured in the `Profiling` segment. The exact testset author pair is excluded from this segment to avoid over-fitting of the classifier. Author pairs for whom citations have been eventuated between 2006 and 2010 are considered for this segment.

`Testing`: The proposed influence scores and the baseline (citation count) are validated using the bibliographic interactions in this segment. An author pair is valid for testing as long as we have individual vector representations for both the authors. Citation exchanges resulted since 2011 are considered for this segment.

6.3 Evaluation and Results

Using the `Training` set, edge representations are constructed for each pair of authors between whom citations are exchanged during this segment. Consequently, relative citation ratios between these author pairs are calculated using Equation (7). The edge representations are then classified with the class labels namely, (NI, SI and HI) using RandomForest. For authors spanning across wide spectrum of ranks and extent of influences, this enables us to capture *what kind of authors cite what kind of authors*. Since we aim at comparing the relative predictive capacity of these representations, we focus less on exact classifier settings and report results achieved by each representation using the same parameters.

On evaluation, precision, recall and f-scores for baseline and influences measures are as shown in the tables (1), (2) and (3). From these results, we can see

that even the resultant representations obtained using baseline citation count performs well enough for the multi-label citation prediction task. However, influence measures I_S and I_D can be clearly seen as better performers almost throughout for each of the aforementioned classes, considering the reported precision-recall values. We observe that, for class NI, almost all the three measures perform equally. This can be attributed to the better accuracy of classifier for non-existent edges between author pairs, irrespective of the weights in consideration. It can also be theorized that certain influence associations may *traverse* from a class to another over a span of time. For example, the citations of the citing author with higher influence of the cited author possibly might get narrowed down (and vice versa) over a period of time. This can happen due to various possible reasons such as a shift in research trends, cultivation of interests in newer fields, etc. Thereby, we also witness recall values for class HI on lower sides for all the three measures.

Table 1: Citation Count

Label	Precision	Recall	F-Score
NI	0.93	0.89	0.91
SI	0.72	0.81	0.76
HI	0.71	0.49	0.58

Table 2: Influence (I_S)

Label	Precision	Recall	F-Score
NI	0.93	0.90	0.91
SI	0.75	0.83	0.79
HI	0.74	0.54	0.62

Table 3: Influence (I_D)

Label	Precision	Recall	F-Score
NI	0.94	0.89	0.91
SI	0.75	0.85	0.80
HI	0.76	0.55	0.64

7 CONCLUSION AND FUTURE WORK

In this paper, we present a model to trace intellectual influences harnessing the structural connectivity within an academic network. Further, we generate author profiling by mapping the latent features into a vector space using the proposed influence scores. We also evaluate effectiveness of the captured author relationships and the resultant author representations by performing experiments for classification tasks, such as citation prediction and the extent of citation. It is observed that results obtained using the suggested influence scores perform better as compared to immediate citation counts. A future direction would be to incorporate types of citations (as mentioned in Section (2.2)) into the current model, possibly using ontological representation of citations [19]. This might help us in knowing and gaining insights on the nature of influences. It would also be interesting to analyze effects of research trends on influences associations.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems 30(1-7), 107–117 (1998)
2. Brooks, T.A.: Evidence of complex citer motivations. Journal of the Association for Information Science and Technology 37(1), 34–36 (1986)
3. Chen, P., Xie, H., Maslov, S., Redner, S.: Finding scientific gems with googles pagerank algorithm. Journal of Informetrics 1(1), 8–15 (2007)

4. Chikhaoui, B., Chiazzaro, M., Wang, S., Sotir, M.: Detecting communities of authority and analyzing their influence in dynamic social networks. ACM Transactions on Intelligent Systems and Technology (TIST) 8(6), 82 (2017)
5. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 855–864. ACM (2016)
6. Jiang, X., Sun, X., Zhuge, H.: Graph-based algorithms for ranking researchers: not all swans are white! Scientometrics 96(3), 743–759 (2013)
7. Kaplan, N.: The norms of citation behavior: Prolegomena to the footnote. Journal of the Association for Information Science and Technology 16(3), 179–184 (1965)
8. Liu, L., Tang, J., Han, J., Yang, S.: Learning influence from heterogeneous social networks. Data Mining and Knowledge Discovery 25(3), 511–544 (2012)
9. Ma, N., Guan, J., Zhao, Y.: Bringing pagerank to the citation analysis. Information Processing & Management 44(2), 800–810 (2008)
10. MacRoberts, M.H., MacRoberts, B.R.: Problems of citation analysis: A critical review. Journal of the American Society for information Science 40(5), 342–349 (1989)
11. Merton, R.K.: The matthew effect in science, ii: Cumulative advantage and the symbolism of intellectual property. isis 79(4), 606–623 (1988)
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
14. Moravcsik, M.J., Murugesan, P.: Some results on the function and quality of citations. Social studies of science 5(1), 86–92 (1975)
15. Nicolaisen, J.: Citation analysis. Annual review of information science and technology 41(1), 609–641 (2007)
16. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: Online learning of social representations. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 701–710. ACM (2014)
17. Pinski, G., Narin, F.: Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. Information processing & management 12(5), 297–312 (1976)
18. Rakoczy, M.E., Bouzeghoub, A., Gancarski, A.L., Wegrzyn-Wolska, K.: Influence in time-dependent citation networks. In: 2018 12th International Conference on Research Challenges in Information Science (RCIS). pp. 1–11. IEEE (2018)
19. Shotton, D.: Cito, the citation typing ontology. In: Journal of biomedical semantics. vol. 1, p. S6. BioMed Central (2010)
20. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 990–998. ACM (2008)
21. Wang, H., Shen, H., Cheng, X., et al.: Scientific credit diffusion: Researcher level or paper level? Scientometrics 109(2), 827–837 (2016)
22. Zhou, Y.B., Lü, L., Li, M.: Quantifying the influence of scientists and their publications: distinguishing between prestige and popularity. New Journal of Physics 14(3), 033033 (2012)
23. Zuckerman, H.: Citation analysis and the complex problem of intellectual influence. Scientometrics 12(5-6), 329–338 (1987)