Measuring the Impact of Topic Drift in Scholarly Networks

Tehmina Amjad[a]
[a]Department of Computer Science and Software Engineering, International Islamic University, Islamabad, Pakistan
tehminaamjad@iiu.edu.pk

Ali Daud[b,a]
[b]Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia adfmohamad@kau.edu.sa

Min Song[c]
[c]Department of Library and Information Science, Yonsei University, Seoul, South Korea
min.song@yonsei.ac.kr

ABSTRACT

With the increase in collaboration among researchers of various disciplines, changing the research topic or working on multiple topics is not an unusual behavior. Several comprehensive efforts have been made for predicting, quantifying, and studying the researcher's impact. The question, that how the change in the field of interest over time or working in more than one topics can influence the scientific impact, remains unanswered. In this research, we study the effect of topic drift on the scientific impact of an author. We apply Author Conference Topic (ACT) model to extract topic distribution of individual authors who are working on multiple topics to compare and analyze with authors who work on a single topic. We analyze the productivity of the authors on the basis of publication count, citation count and h-index. We find that authors who stick to one topic, produce a higher impact and gain more attention. To further strengthen our results we gather the h-index of top-ranked authors working on one topic and top-ranked authors working on multiple topics and examine whether there are similar trends in their progress. The results show an evidence of significant impact of topic drift on career choices of researchers.

CCS CONCEPTS

• **Information systems** → **Information retrieval** → document topic models

KEYWORDS

Academic Social networks, Scientific Impact, Ranking of Authors, Topic-based Ranking, Change of Topic.

ACM Reference format:

T. Amjad, A. Daud, and M. Song. 2018. Measuring the Impact of Topic Drift in Scholarly Networks for WWW 2018 Proceedings and Companion Submissions. In *The 2018 Web Conference Companion (WWW 2018), April 23-27, 2018, Lyon, France*, ACM, New York, NY, XX pages. DOI: https://doi.org/10.1145/3184558.3186358

WWW '18 Companion April 23-27, 2018, Lyon, France.

ACM ISBN 978-1-4503-5640-4/18/04.
DOI: https://doi.org/10.1145/3184558.3186358

1 INTRODUCTION

Scientists are always interested in examining the truth and have critical questioning abilities. With the help of such abilities, they produce new publications in their field of interest. At different phases of their career trajectory, the scientists are engaged in academic and research activities in a way that is beneficial to their environment and are concerned about their own career evolution and progress as well [1]. The career progression that may include the academic promotion or obtaining the research grants requires a stable and non-interruptive stream of publications. It can be achieved with the help of sound contributions to an established research plan. A persistent research profile help scholars to remain constructive but perhaps it weakens the likelihood of originality [2]. Pioneering perceptions usually arise from encountering new challenges and attempting new topics or integrating them with existing research agenda. The researchers may try to explore new challenges in an urge of progress and evolution of a career. Apart from that, the nature of across discipline collaborations has made the change in the field of interest unavoidable phenomena in career trajectory of many scholars.

Despite its importance on a scientific career, productivity, research honorarium and hiring procedures, the impact of topic drift in career trajectory of a researcher remains unexplored. The researchers can achieve different levels of expertise in all their fields of interest. This variation in the field of interest can influence the performance of the researchers and academicians. The effect of this topic drift has not been yet measured in literature.

The primary motivation of the present study is to find the impact of the shift of topic on the career of a scholar, as several scholars are interested in working in more than one topics and some have a tendency to change their topic of interest over time. The findings will help those researchers who are keen on gaining and maintaining a high standing in their scholarly networks by knowing that they shall concentrate more on a single topic or attempt to be more diverse. Several studies have been conducted to measure the impact of researchers. It is known that received citations can play a very significant role in measuring the impact of a researcher [3], [4]. Several studies have proposed methods for topic-based ranking of authors by calculating their impact in all the fields they have worked [5], [6]. Various efforts have measured the effect of publication venue on the ranking of authors [7]. In this paper, we study the effect of working on multiple topics on the impact of authors. For the said purpose, we perform experimentation on a dataset retrieved from the field of Information Retrieval (IR) and we analyze the progress of

authors working in single or multiple sub-topics of IR. Learning that a researcher can improve the chances of receiving citations by working in more than one topic or not can have many cognitive benefits. The results of the study show that working on more than one topics cannot significantly improve the scientific impact of a researcher.

The rest of the paper provides is organized as follows: related work to the present study is provided in section 2; the problem statement is denoted in section 2; details of dataset selection are given in section 3; section 4 describes the proposed method; section 5 includes discussion on obtained results; section 6 concludes the study and provides future directions.

2 RELATED WORK

2.1 Author Ranking Methods

From the literature, we have identified a number of methods for ranking of authors which consider several features as weights. Author Rank algorithm [8] calculates author significance within a co-authorship network considering collaboration intensity. Author's impact was quantified by incorporating both citation and co-authorship graph property and weighted versions of PageRank were presented in [9], [10]. Six different variants of the bibliographic network were studied and compared and as a result, the hybrid networks were recommended for studying research interactions and collaborations [11]. While some studies have evaluated heterogeneous scholarly networks such as; a combination of combining authorship, journal-ship, and publication citation networks [12], citation networks of papers, authors, associations and publishing venues [13], authors, publications, and publishing venues along with the effect of the topic [5] were simultaneously considered. The mutual influence of authors on each other was studied and it was established that ranking of an author is influenced by the standing of their co-authors, highly ranked co-authors have more influence on junior co-authors and vice versa [3], [14]–[16]. The effect of collaboration with a top-notch researcher on the career of a junior was analyzed and significant findings were revealed [17]. Q-Rank was presented for ranking of authors [18]. It computes the preliminary rank score of a researcher depending upon qualities of papers as well as the participation of the researcher to those papers. It also deals with the situations in which the papers are published by only one author.

The citations based methods, as well as PageRank based methods, were used to predict the award-winning researchers from approximately two million textual metadata records on computer science papers and impact of citations was studied [19]. Codd Award and Turing Award winners were chosen to evaluate the results and PageRank based methods perform better when relative ranks of awardees were considered.

2.2 Temporal Methods

The temporal dimension is also an important factor and can play an essential role when incorporated as a ranking criterion [20]. A time aware method was presented that considers the time of publication and time of the citation. The method assigned weights to the citations by considering that whether and when two authors have collaborated with each other [21]. Time rank method considered the time of citations and gives more weight to authors that are cited by highly ranked authors [22]. The method showed that timing of recognition of an author is significant for academic tasks like finding rising stars.

Heterogeneous factors including citations, authors, publication venues and the publication time information were considered for ranking of scientific articles [23]. A time aware method for ranking of authors, TAPRank, studies heterogeneous networks for finding author's impact considering the time of publication [24].

2.3 Topic-based Methods

Topic-based methods were used to identify researchers with given expertise for different time periods [25]. Some topic based methods have been proposed [6], [26], [27] yet these variants do not consider the heterogeneity of an academic system. In a scientific domain, the problem of topic evolution was studied for finding key domain advancements and for assisting the knowledge transfer within and across domains. It was identified that the growth of the main topic normally trails a pattern from adjusting status to mature status in an evolving process [28]. Analysis of diffusion of topics was studied recently for a weighted citation network and a framework for detection of topic evolution was presented [29].

A topical inheritance model, between cited and citing papers, was used to study the strength of citation influence and identification of paper topics. However, these methods that involve the impact of topic for ranking of authors along with some other criterion do not measure the impact of the change of topic by a researcher during the research career. A recent study has performed a large scale analysis on the bibliographic dataset and it was found that changes in research interest follow a reproducible pattern in form of exponential distribution [30]. In a quantitative way nature of changes in research interest of researchers has been studied in their careers.

2 PROBLEM STATEMENT

The authors have potential to work in more than one filed and they may have different scoring in different fields of their choice. The impact of drift of topic needs to be analyzed to find that authors who are working in multiple topics were able to produce more impact or authors who are working in the single topic are successful in their career. The analysis will help the researchers to decide that they shall focus on a single topic or shall try working on more than one topics to attain higher impact during their research career.

3 DATASET DESCRIPTION

We chose the field of IR as our testbed, and the papers and their cited references were extracted from Web of Science for the period 1956 to 2014. The Search strategies were based on the following terms: Information Retrieval, Query Language, Query Languages, Information Storage and Retrieval, Query Processing, Database Query, Database Queries, Document Retrieval, Data Retrieval, Image Retrieval, Text Retrieval, Content-Based Retrieval, Content-Based Retrieval, and Relevance Feedback [5]. In total, we have collected 20,359 papers with 44,770 distinct authors, 558,498 citations relationships, and 3,270 unique journals.

The title and abstract fields are used as the text corpus for extracting topics.

4 METHODOLOGY

For the current study, we need to classify authors into two main categories, i.e., (1) authors working in the single topic and (2) authors working on multiple topics. The brief methodology is given in Figure 1. For analysis, we only consider the publications that receive one or more citations. We have calculated the cumulative number of citations for all the publications made by the author as a performance parameter. The analysis was based on results extracted in five different phases. First of all, extraction and classification of authors according to their topics was performed. In this phase, single topic authors (STA) and multiple topic authors (MTA) were separated by using ACT model [31]. It considers the probability distribution of all authors over extracted topics (i.e., the probability of a topic for a given author: P (t|a)). The author's probability distribution for a given topic was calculated by model (i.e., P (a|t) =P (a) P (t|a)/P (t)). ACT model calculated the probability distribution of author and topic simultaneously:

- The probability of a topic for a given author P(t|a) which infers research interest of the given author, the probability of an author for a given topic P(a|t) which can derive the most productive authors for the given topic;
- The probability of a topic for a given document P(t|d) which infers the topic distribution of the given document, the probability of a document for a given topic P(d|t) which can derive the most related papers to the given topic

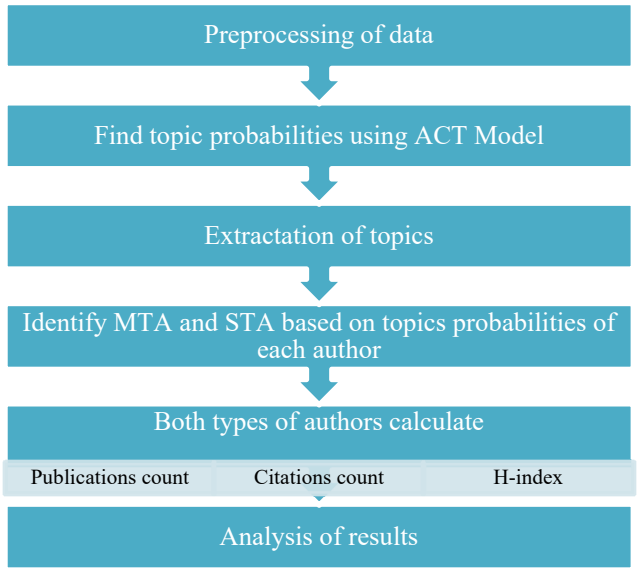

Figure 1: Brief methodology of the conducted study

Using ACT model we calculated the probability of a topic for a given author and the probability of a topic for a given document. In the second phase, the preprocessing of data was performed. The authors who work on one topic or on multiple topics were identified after initial preprocessing. It was obtained by scanning of topic probabilities of all authors in ten different topics from IR. First, we find out the highest value of topic for an author. Out of all ten topics if there exists only one highest value then it shows that the author is working only on one topic and we have categorized it as STA. If an author has higher values for more than one topic that means that this author is working on more than one topics and we categorize them as MTAs. We also identified the topics in which these authors are working. In the next phase, publication count of authors with respect to the topic was calculated in which publications of STAs and MTAs were calculated and stored in the descending order. Similarly, citations count of authors with respect to the topic were also calculated and stored for both types of authors. Finally, we extracted the h-index for the authors from the AMiner website [32]. By the end of this process, we came up with two lists of authors, who worked in a single topic and who worked in multiple topics along with their number of publications, citations, and h-index. Table 1 represents the pseudocode followed for obtaining the desired data regarding authors.

Table 1: The pseudocode followed for obtaining results to be analyzed

```
Perform initial preprocessing of data
Apply ACT model to find topic probabilities of authors
Extract 10 Topics from IR
Analyze topic probabilities of all authors for each topic
Based on topic probability classify authors as
        MTA
        STA
For each MTA and STA
        Calculate Citation Count
        Calculate Publication Count
        Retrieve H-index from AMiner
```

5 RESULTS AND DISCUSSION

In this section, we briefly discuss the results obtained for authors working in single topic and authors working in multiple topics individually and then we compare them to conclude our work. Table 2 shows the representative words of ten topics identified by the ACT. First of all, we have arranged the authors with respect to their decreasing citations count for both topics and compare both types of authors. Figure 2 represents the comparison of citation count between MTAs and STAs. First five authors of both categories have the same citation count (106). However, amongst authors from 6-11, it was found that STAs have a greater number of citations as compared to MTAs. Overall, the total citations across top 15 STAs and MTAs are 840 and 763 respectively indicating that STAs receive more citations as compared to MTAs. On average, top 15 authors of single and multiple topics receive

56 and 50 citations respectively. The trend line shows that higher trend was obtained by the STAs as compared to MTAs. Figure 3 compares the publication count of the authors working on single topic and authors working on multiple topics. It is generally assumed that people working on multiple topics can produce more publications as they are more diverse and can gain attention from more diverse people. In Figure 3, we arranged the authors with respect to their decreasing number of publications. It depicts that authors working on the single topic were able to produce more publications as compared to authors working on multiple topics. Thus showing that focusing on a single topic can result in more publications.

Table 2: Representative words of ten extracted topics

	Representative words for ten topics from IR via ACT
Topic 1	Image, retrieval, content, color, visual, video, shape
Topic 2	Retrieval, information, language, document, query, relevance, text
Topic 3	Query, data, database, queries, xml, efficient, search
Topic 4	Query, data, processing, networks, mobile, sensor, distributed
Topic 5	Information, search, web, user, library, study, analysis
Topic 6	Web, information, semantic, ontology, management, framework, analysis
Topic 7	Retrieval, information, system, data, database, object, relational
Topic 8	Retrieval, information, data, image, relevance, indexing, ranking,
Topic 9	Data, medical, information, database, clinical, health, analysis
Topic 10	Retrieval, information, document, fuzzy, web, image, knowledge

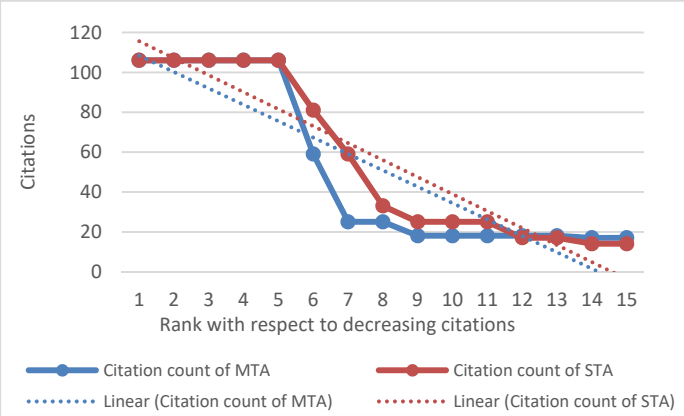

Figure 2: Comparison of citations count of single and multiple topic authors

Moreover, the total number of publications by top 15 authors working on single topics is 358, whereas, the total number of publication by top 15 authors working on multiple topics is 307. This finding is not in alliance with the general assumption, that is, one can produce more publications by working on more topics instead of focusing on a single topic. The trend line clearly shows that STAs were more productive as compared to MTAs. The linear trend lines of citations count and publication count both show a similar pattern depicting seniority of authors working in the single topic over the authors working in multiple topics.

The values of publication count and citation count are dependent upon the publications present in our selected dataset and hence are limited to the size of the dataset.

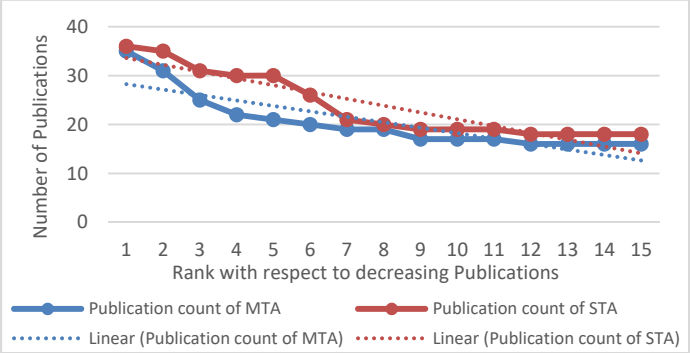

Figure 3: Comparison of publications count of single and multiple topic authors

In literature, the citations are considered as a parameter for measuring the performance of the authors. To cross-check the validity of the trend found from the dataset, we have taken the top 15 most cited authors along with their number of publications, and we have taken their h-index values from the AMiner[1]website. Unlike the citations and publications count, the h-index values gathered from AMiner are not dependent on the statistics of the dataset and hence they give a wider picture of the standing of authors. Table 3 shows the top 15 most cited STAs and we can see their h-index and publications. Similarly, Table 4 shows the top 15 most cited MTAs along with their h-index and publications. For both cases, we took fifteen authors who are sorted on the basis of their decreasing citation count (there are 5 authors in a dataset with highest citation count). These tables show that for the majority of authors, with an increase in publication count, the h-index is also increasing. It is noticed that top 15 authors of both categories with the almost similar number of publications have very distant values for h-index, while the highest value of h-index was achieved by an author of the single topic category. Figure 4 plots the publications from the dataset for the top 15 most cited authors from both topics. The trend that can be seen in this figure

[1] https://www.aminer.cn/

is also in accordance that was identified from Figure 3. Moreover, the total number of publications by top 15 STAs (with respect to citations count) is 92, whereas, the total number of publication by top 15 MTAs (with respect to citations count) is 62 which is a big difference between the two types of authors. Figure 5 shows that the trend of h-index found from AMiner is also going in accordance with the results obtained from the dataset under consideration and it further affirms our results. Just like Figure 4, the h-index values are for the top 15 highly cited authors for both topics. Thus it can be stated that working on multiple topics do not guarantee that the researchers will be able to produce more publications and can gain more attention from the scientific community via citations. Giving more attention and focusing a single topic can be a reason for a modest increase in scientific performance.

Table 3: The citation count, publication count and h-index for top 15 most cited STA

Author name	citation count	publication count	h-index
Lee D	106	3	30
Sawahney H	106	1	12
Steelee D	106	1	7
Gorkani M	106	1	3
Yanker P	106	1	3
Salton G	81	30	48
Grey JE	59	2	3
Zhou XS	33	6	28
Del Bimbo A	25	17	36
Pala P	25	15	21
Colombo C	25	3	1
Kohonen T	17	3	37
Honkela T	17	2	21
Zhang C	14	5	1
DeWitt D	14	2	89

Table 4: The citation count, publication count and h-index for top 15 most cited MTA

Author name	citation count	publication count	h-index
PetKovic D	106	2	22
Niblack W	106	1	12
Dom B	106	2	5
Huang Q	106	2	1
Hafner J	106	1	1
Mehrotra R	59	3	1
Ceri S	25	13	64
Bonifati A	18	4	24
Geissbuhler A	18	7	13
Muller H	18	12	29
Rosset A	18	2	4
Garcia A	18	3	1
Vallee JP	18	2	10
Valencia A	17	3	70
Kaski S	17	5	37

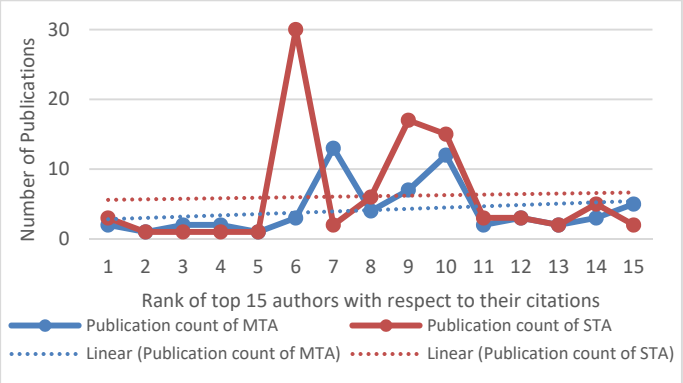

Figure 4: Comparison of publications count of single and multiple topic authors of top 15 most cited authors

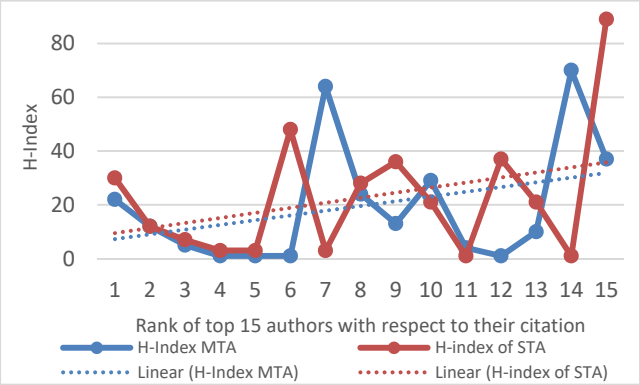

Figure 5: Comparison of H-index of single and multiple topic authors

6 CONCLUSIONS AND FUTURE WORK

Authors have a tendency to change their topic of interest over time or they may work on multiple topics at the same time. Researchers may not usually be an expert in all the fields they are interested in. They can be a top-notch researcher in one topic and a mediocre in another topic at the same time. In this study, we find the effect of change of topic and working on multiple topics on the ranking of authors. From the trends shown in the figures in results and discussion section, we find that authors who work on a single topic tend to be more productive and they have a high publication count, a high citation count, and a greater h-index value. While the authors who work on multiple topics have comparatively less number of publications, citations, and h-index.

For dataset selection, we limited the field of interest to information retrieval. As a result of this dataset selection activity, the findings are biased towards scientists working in one or multiple subfields of IR. Yet, this limitation also indicates towards new openings for further explorations. In fact, more wide-ranging datasets can provide new and richer understandings towards the impact of topic drift across different disciplines. Further investigations regarding the impact of topic drift across more

generic specialties should also be discovered as it could uncover more significant results. Taken together our results offer the first systematic empirical evidence on how topic drift affect scientific performance and impact.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A3A2046711). We also acknowledge Ms. Mahrukh Malik for her support at the initial stages of experimentation.

REFERENCES

[1] H. Ankarali, O. Pasin, and S. Ankarali, "A New Index for Evaluating Academic Performance: Hos-index," *Int. J. Comput. Sci. Issues IJCSI*, vol. 14, no. 3, p. 78, 2017.

[2] J. G. Foster, A. Rzhetsky, and J. A. Evans, "Tradition and innovation in scientists' research strategies," *Am. Sociol. Rev.*, vol. 80, no. 5, pp. 875–908, 2015.

[3] T. Amjad, A. Daud, D. Che, and A. Akram, "MuICE: Mutual Influence and Citation Exclusivity Author Rank," *Inf. Process. Manag.*, 2015.

[4] D. Bouyssou and T. Marchant, "Ranking authors using fractional counting of citations: An axiomatic approach," *J. Informetr.*, vol. 10, no. 1, pp. 183–199, 2016.

[5] T. Amjad, Y. Ding, A. Daud, J. Xu, and V. Malic, "Topic-based heterogeneous rank," *Scientometrics*, vol. 104, no. 1, pp. 313–334, 2015.

[6] T. Amjad and A. Daud, "Indexing of authors according to their domain of expertise," *Malays. J. Libr. Inf. Sci.*, vol. 22, no. 1, pp. 69–82, 2017.

[7] D. Peiris and R. Weerasinghe, "Citation network-based framework for ranking academic Publications and venues," in *Advances in ICT for Emerging Regions (ICTer), 2015 Fifteenth International Conference on*, 2015, pp. 146–151.

[8] X. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel, "Co-authorship networks in the digital library research community," *Inf. Process. Manag.*, vol. 41, no. 6, pp. 1462–1480, 2005.

[9] E. Yan and Y. Ding, "Discovering author impact: A PageRank perspective," *Inf. Process. Manag.*, vol. 47, no. 1, pp. 125–134, 2011.

[10] D. Fiala, F. Rousselot, and K. Ježek, "PageRank for bibliographic networks," *Scientometrics*, vol. 76, no. 1, pp. 135–158, 2008.

[11] E. Yan and Y. Ding, "Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other," *J. Assoc. Inf. Sci. Technol.*, vol. 63, no. 7, pp. 1313–1326, 2012.

[12] E. Yan, Y. Ding, and C. R. Sugimoto, "P-Rank: An indicator measuring prestige in heterogeneous scholarly networks," *J. Am. Soc. Inf. Sci. Technol.*, vol. 62, no. 3, pp. 467–477, 2011.

[13] Z. Yang, L. Hong, and B. D. Davison, "Topic-driven multi-type citation network analysis," in *Adaptivity, Personalization and Fusion of Heterogeneous Information*, 2010, pp. 24–31.

[14] T. Amjad, A. Daud, A. Akram, and F. Muhammed, "Impact of mutual influence while ranking authors in a co-authorship network," *Kuwait J. Sci.*, vol. 43, no. 3, 2016.

[15] X.-L. Li, C. S. Foo, K. L. Tew, and S.-K. Ng, "Searching for rising stars in bibliography networks," in *Database Systems for Advanced Applications*, 2009, pp. 288–292.

[16] A. Daud *et al.*, "Finding Rising Stars in Co-Author Networks via Weighted Mutual Influence," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 33–41.

[17] T. Amjad *et al.*, "Standing on the shoulders of giants," *J. Informetr.*, vol. 11, no. 1, pp. 307–323, 2017.

[18] L. Li, X. Wang, Q. Zhang, P. Lei, M. Ma, and X. Chen, "A quick and effective method for ranking authors in academic social network," in *Multimedia and Ubiquitous Engineering*, Springer, 2014, pp. 179–185.

[19] D. Fiala and G. Tutoky, "PageRank-based prediction of award-winning researchers and the impact of citations," *J. Informetr.*, vol. 11, no. 4, pp. 1044–1068, 2017.

[20] T. Amjad, A. Daud, and N. R. Aljohani, "Ranking authors in academic social networks: a survey," *Libr. Hi Tech*, vol. 36, no. 1, pp. 97–128, 2018.

[21] D. Fiala, "Time-aware PageRank for bibliographic networks," *J. Informetr.*, vol. 6, no. 3, pp. 370–388, 2012.

[22] M. Franceschet and G. Colavizza, "TimeRank: A dynamic approach to rate scholars using citations," *J. Informetr.*, vol. 11, no. 4, pp. 1128–1141, 2017.

[23] Y. Wang, Y. Tong, and M. Zeng, "Ranking Scientific Articles by Exploiting Citations, Authors, Journals, and Time Information.," 2013.

[24] X. Kong, J. Zhou, J. Zhang, W. Wang, and F. Xia, "TAPRank: A Time-Aware Author Ranking Method in Heterogeneous Networks," in *Smart City/SocialCom/SustainCom (SmartCity), 2015 IEEE International Conference on*, 2015, pp. 242–246.

[25] A. Daud, J. Li, L. Zhou, and F. Muhammad, "Temporal expert finding through generalized time topic modeling," *Knowl.-Based Syst.*, vol. 23, no. 6, pp. 615–625, 2010.

[26] Z. Yang, J. Tang, J. Zhang, J. Li, and B. Gao, "Topic-level random walk through probabilistic model," in *Advances in Data and Web Management*, Springer, 2009, pp. 162–173.

[27] T. A. Olowookere, B. O. Eke, and L. U. Oghenekaro, "A Topic Modelling-Based Framework for Mining Digital Library's Text Documents," 2015.

[28] B. Chen, S. Tsutsui, Y. Ding, and F. Ma, "Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval," *J. Informetr.*, vol. 11, no. 4, pp. 1175–1189, 2017.

[29] M. Kim, I. Baek, and M. Song, "Topic diffusion analysis of a weighted citation network in biomedical literature," *J. Assoc. Inf. Sci. Technol.*

[30] T. Jia, D. Wang, and B. K. Szymanski, "Quantifying patterns of research-interest evolution," *Nat. Hum. Behav.*, vol. 1, p. 78, 2017.

[31] J. Tang, R. Jin, and J. Zhang, "A topic modeling approach and its integration into the random walk framework for academic search," 2008, pp. 1055–1060.

[32] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," 2008, pp. 990–998.