Scientometrics (2018) 116:675–687
https://doi.org/10.1007/s11192-018-2787-6

A quantitative exploration on reasons for citing articles from the perspective of cited authors

Binglu Wang[1] · Yi Bu[2] · Yang Xu[1,3]

Received: 25 April 2017 / Published online: 2 June 2018

Abstract Citation is regarded as one of the "norms of science" (Merton in Am Sociol Rev 22(6):635–659, 1957) and is deeply researched by the field of scientometrics. The motivations authors have for citing one another are considered significant and have been the subject of extensive qualitative research such as content analysis, questionnaires, and interviews of citing authors. However, the existing qualitative studies have covered a limited number of samples. To expand the dataset, this paper proposes a quantitative method applied to detecting citation reasons from the angle of citation networks and the attributes of cited authors, including their publication count (the number of single-authored publications, collaborative and first-authored publications as well as collaborative but non-first-authored publications, and number of whole publications), citation count, research topic interests, and gender. By applying the Exponential Random Graph Models (ERGMs), the current study revealed that authors in the field of information retrieval tend to cite those with more single-authored, collaborative and first-authored, and collaborative but not first-authored publications. Besides, in this field, the number of publications, similar topical domains, and same gender are proven to be significantly favorable in selecting references in our experiment.

Keywords Citing behavior · Exponential Random Graph Models (ERGMs) · Citation network · Bibliometrics · Scientometrics

✉ Yang Xu
yang.xu@pku.edu.cn

[1] Department of Information Management, Peking University, Beijing, China

[2] Center for Complex Networks and Systems Research, School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA

[3] Ocean Strategy Research Center, Peking University, Beijing, China

Introduction

The emphasis on citing stemmed from 1920s, when Gross (1927) applied citation count to evaluate the impact of research. Citing is an important behavior in science, commonly regarded as a key part of research discipline, the main communication pathway between scholars (Garfield 1972), and one of the hypothesized "norms of science" (Merton 1957). Additionally, citing behavior has become a significant factor in maintaining knowledge accumulation and research development (Cronin 1984). Speaking in detail, it not only inherits and transcends the knowledge basis diachronically, but also expands and enriches our discoveries synchronously.

Citing has contributed enormously in information retrieval. Garfield (1955) was the first to propose that citing instead of traditional headwords could be applied as tags to document retrieval. Indexing by citations provides the following advantages: (1) the indices are stable, professional, and objective; (2) there is no limitation of subjects, and of much convenience for interdisciplinary research retrieval; (3) citing retrieval language is appropriate and feasible for automatic machine processing. Especially in the digital era, the combination of citing retrieval language and hyperlink technologies has greatly facilitated information and resources retrieval.

However, the citing motivations are sometimes far from our expectation. A paper can be cited simply because of its high research quality and reputation significance but relevance. Additionally, the citing motivation is also affected by other factors, such as exaggerating one's research ability, adding a large quantity to citation to show one's professionalism, and citing unrelated papers blindly (Cano 1989; Garfield 1998). Thus, the motivations for citation are often murky, which complicates research on the subject. Qualitative methods like interviews have been widely used in the previous research about citing behavior (Lipetz 1965; Brooks 1986). With the limitations of manual processing, the number of samples used in these studies was rather small, which might cause certain analytical biases and negatively affect the accuracy of results. For potential biases of limited sampling, a quantitative exploration on a large-scale dataset is applied in this paper to help us better understand reasons for citing, thereby increasing the accuracy and persuasiveness of the conclusion.

Moreover, previous research about citing reasons tended to focus on the authors' *stated* (*claimed*) reasons for making a citation (Garfield 1964; Gilbert 1977; Carpenter and Narin 1981; Prabha 1983; Cole and Singer 1991; Baldi 1998; Van Raan et al. 2003; Stack 2004; Erikon and Erlandson 2014), overlooking potential subconscious reasons for citations. Different from some qualitative methods such as interview that cannot deeply reveal the subconscious reasons for citations, the current paper proposes a quantitative exploration that directly shows whether the considered factors have effects on authors' citing other researchers' papers.

Additionally, most research digs out the pattern from the angle of citing authors rather than cited ones. Nevertheless, the citation networks of cited authors also provide many clues about why they are cited (Li et al. 2016), thus encouraging us to consider citing reasons from cited authors' perspectives. The main research question in the current study is: What factors related to cited authors' bibliometrical and gender attributes relate to citation behavior? To address this research question, we select several important indicators related to the cited authors, including the number of publications, the number of citations, research topic, and gender of cited authors, to weigh their influence on the formation of citation network by employing the Exponential Random Graph Model (ERGM) (Robin et al. 2007a). The main and homophily effects of these indicators are detected simultaneously to explore the actual citing reasons from the cited authors' perspective.

This article is outlined as follows. In the following section, we will illustrate the related work about citing reason studies. The dataset we use as well as the indicators and model we employ will be discussed in detail in the third part. The results and discussion are illustrated in the fourth part before our conclusions.

Related studies

From the 1960s onwards, scholars gradually noticed the research value inherent in citation patterns (Garfield 1965; Lipetz 1965). When gathered together, following up on these studies, researchers began to model the citing and cited literatures as mammoth citation networks that they used to analyze the fields' relationship, scientific communication, and knowledge evolution (Merton 1968; Cronin 1984).

As mentioned above, citing behavior is an important phenomenon in knowledge evolution, and uncovering the citing reasons becomes important for us to understand this process (Garfield 1965). Garfield (1964) proposed 15 types of motivations for citation by analyzing citing locations, contents, and forms: (1) paying homage to pioneers; (2) giving credit for related work (homage to peers); (3) identifying methodology or equipment; (4) providing background reading; (5) correcting one's own work; (6) correcting the work of others; (7) criticizing previous work; (8) substantiating claims; (9) alerting to forthcoming work; (10) providing leads to poorly disseminated, poorly indexed, or uncited work; (11) authenticating data and classes of fact or physical constants; (12) identifying original publications in which an idea or concept was discussed; (13) identifying original publication or other work describing an eponymic concept or term; (14) disclaiming work or ideas of others (negative claims); and (15) disputing priority claims of others (negative homage). Based on previous research (Lipetz 1965; Weinstock 1971), Brooks (1986) divided the citing motivations into seven types, including currency scale, negative credit, operational information, persuasiveness, positive credit, reader alert, and social consensus.

Due to the complexity of citing motivation, it is difficult to understand the authors' citing behavior simply by the content of literature or citation analysis (Gilbert 1977). Therefore, many scholars have applied interview methods to remind authors of their thoughts when they wrote a paper (Prabha 1983). Compared with context analysis, interviews can give insights into authors' inner thoughts and their citing reasons. Liu (1993), for example, collected 415 questionnaires from authors who had published articles on *Chinese Physics* during 1981–1987, aiming to construct a theoretical model about citing motivation and the potential relationships among different factors. Cases and Higgins (2000) applied the interview method to survey highly-cited authors in the field of media studies who published articles between 1995 and 1997. Based on their results, they divided most citing motivations into three types: (1) for currency, reputation, and concept explanation; (2) improving the authority of the research; and (3) regarding the reference as a value.

Additionally, citing reasons for specific groups and topics vary from one another. Kapseon (2004), for instance, has anatomized the particular literature citing motivation for social scientists in South Korea by bibliographic coupling (Zhao and Strotmann 2008, 2014) and demonstrated that citing English-written literature seems much more professional for those non-English-native countries. Hendley (2012) focused on undergraduates majoring in history, politics, and sociology by interviewing these students and argued that they tend to refer to related books and journals via the Internet with major differences between each subject.

While most scholars have considered citing motivations from the perspective of the citing authors, there are still many contributions focusing on the cited side as well as the dynamics between citing and cited sides. The pattern of citation from both citing and cited perspectives can be connected once rescaled by the growth of publications and citations, which indicates that citing and cited behavior are indeed "two sides of the same coin" (Yin and Wang 2017). In this scenario, people also explore the citation reason from the cited papers other than citing authors. By constructing two citation-based indicators that generated from a critical sub-network, Li et al. (2016) provided a framework to weigh a paper's latent referential value. Their result shows that papers with high quality, considerable quantity, and hotpots are easier to get cited. Besides, articles that have a higher vitality performance in recent years regardless of their total citation tend to be cited more likely (Wang et al. 2017). More properties of cited papers like the structure of citation networks (Chen 2012) are also discussed in previous research. Similarly, based on the citation network, this paper applied a quantitative approach of network structure, ERGM (Exponential Random Graph Model) (Robin et al. 2007a, b), to detecting citing reasons from the cited authors' perspective. Compared with other citated side analysis method, ERGM takes both papers' properties (i.e., nodes' attributes) and the structure of citation network into consideration, which not only captures local features but also accounts for global characteristic at the network level.

Above all, citing behaviors have been found to be influenced by various factors. Here we categorized the potential factors by five different types as follows:

(1) Related research topics. Authors learn a lot from previous research especially when there are many similarities in terms of theories, experiments, and methodologies (Cases and Higgins 2000). Duncan (1981), for instance, have come up with 26 relations between cited and citing documents and proved the high similarity between them. Additionally, Chubin and Moitra (1975) emphasized the importance of citation as describing relevant studies and providing historical backgrounds.

(2) Accessibility of literatures. Availability accords with the principle of least effort (Bornmann and Daniel 2008), for the reason that physical accessibility (Soper 1976), publishing media (Silverman 1985), and free online availability of publications (Lawrence 2001) have an impact on citing possibility. Marx and Bornmann (2015) demonstrated that the average citation rate is greatly influenced by the extent to which the papers (cited as references) are accessible as linked database records.

(3) Reputations of journals and authors. Seeking and following authorities is consistent with the Matthew effect, accumulative advantage, and the saying that success breeds success (Cozzens 1985). Many studies also utilized the number of citations of peer reviewed papers to measure the impact of the scholar's work, where high quality work may trigger more responses such as citations from others (Van Raan et al. 2003).

(4) Social alignment like ideology and cultural background. Scientific tradition and historical preconceptions influence the authors' choices, especially citing behavior (Erikon and Erlandson 2014); culture barriers also affect citation probability (Carpenter and Narin 1981).

(5) Individual factors of the citing author. Citing motivations have been found to be impacted by individual research value, citing tendency, citing habits, and citing strategy (Garfield 1998). Sandstrom et al. (2005) have shown that citations are affected by social networks, where authors tend to cite from those they are personally acquainted with. Furthermore, gender is also a vital factor, where men tend to receive substantially more citations than women, indicating a potential gender bias for authors' citation behaviors (Cole and Singer 1991; Baldi 1998; Stack 2004).

Another issue related to the current research is the methodology adoption. Qualitative methods have been widely applied to detect the authors' citing motivations before, but they consume enormous manpower, material, and financial resources, and thus make large samples impossible to handle. In order to solve this problem and analyze the citing reason, this paper employs a quantitative method (ERGM), in which main and homophily effects are both considered at the same time. Additionally, we construct the citation network to understand citing reasons from the perspective of cited authors, which not only emphasizes the relationship between cited and citing authors, but also increases the credibility of the results. Specifically, we utilized authors' bibliographic metadata like the number of publications, the number of citations, gender, as well as most frequently used research topic of cited authors from a relatively massive data set. Corresponding to the five summarized factors we proposed, the number of publications and citations are selected to indicate the reputation of authors (Van Raan et al. 2003). Topic domain demonstrates research topical relatedness, while gender can represent the individual factor. Besides, accessibility and reachability of literatures reflect the attribution of citing authors, while we focus on cited authors in this paper. Based on the large-scale and structural data, we adopted Exponential Random Graph Model (ERGM) in this paper, aiming to dig out the potential citation reasons (motivations) behind citation network by examining main and homophily effects in the same model from the cited authors' perspective.

Methodology

Data

The data set we used are the same as that in Zhang et al. (2018). We retrieve the data set from Web of Science (WoS) database with the following query terms: information retrieval, information storage and retrieval, query processing, document retrieval, data retrieval, image retrieval, text retrieval, content based retrieval, content-based retrieval, database query, database queries, query language, query languages, and relevance feedback. 20,359 papers in the field of information retrieval published between 1956 and 2014 and their 59,162 authors and 558,498 references are contained. To disambiguate the authors' names, we employed a simple two-step matching procedure based on their full names and affiliations (Yu et al. 2014). After applying this method we identified 44,770 distinct authors in the dataset.

Indicators

Zhang et al. (2018) employed several indicators to understand whether and to what extent they would affect the formation of author collaboration network; these indicators include:

number of single-authored publications, collaborative first-authored publications, collaborative but non-first-authored publications; number of citations; most frequently-used research topic; and gender. We here follow Zhang et al. (2018)'s framework and to use the same indicators on cited authors as those used by their study. Previous studies have shown that the productivity (commonly measured by the number of publications) (Wuchty et al. 2007), popularity (commonly measured by the number of citations) (Ding and Cronin 2011), and research topic similarity might have several effects on the citation network formation (Cases and Higgins 2000); hence, we decided to quantitatively investigate whether these indicators could have main and/or homophily effects on the formation of the citation network. Additionally, the publications are classified into three types—single-authored, collaborative and first-authored, as well as collaborative but non-first-authored number of publications—in the current study to see their main effects. We included gender information in the model to evaluate whether the authors' gender, which could have any significant effects on citing behavior, as suggested by previous work (McDowell and Smith 1992), has significant effects on citing behavior. Here are the indicators involved in this study:

(1) The number of single-authored publications one author has published;
(2) The number of collaborative first-authored publications one author has published;
(3) The number of collaborative non-first-authored publications one author has published;
(4) The number of citations one author has received;
(5) The most frequently used research topic; and
(6) The gender of one author.

For (5) above, we employ the Author-Conference-Topic (ACT) model proposed by Tang et al. (2008), in which the research topic distribution of each author is extracted as a vector and the similarity between topic distributions (vectors) of two authors could be compared by applying cosine similarity. We extract the same five topics (i.e. the components in a topic distribution vector) used by Zhang et al. (2018). We then selected as each author's "core research interest" the topic with the highest weight in his/her distribution. If there are two or more topics with the same highest weight, we randomly select one of them as the representative of his/her "core research interest". As for (6), we manually search gender information on the website of these authors.

Hypotheses

We will test the hypotheses that the formation of the citation network in the field of information retrieval could be influenced by main and homophily effects from the following factors of cited authors:

H1: The number of single-authored publications;
H2: The number of collaborative and first-authored publications;
H3: The number of collaborative but non-first-authored publications;
H4: The number of citations;
H5: The most frequently used research topic; and
H6: The gender

Citation network

As pointed out by Zhao and Strotmann (2008), the most productive authors are often regarded as active and representative of the whole field of development. As a result, we selected the most productive 500 authors in our dataset in order to make the network denser. However, since the authors ranked from 500th to 633rd have published the same number of papers, we have to include all of them. For the citation network formation, if author A has cited the author B at least once, then there will be a tie from A to B; the strength of the tie from A to B equals to the number of citations B has received from A.

Exponential Random Graph Model (ERGM)

We applied the Exponential Random Graph Model (ERGM) to model the citation network and the attributes of the nodes (authors). The ERGM regards as the explanations of network any network statistics and the nodes' attributes in the observed network (Wasserman and Pattison 1996; Handcock et al. 2003; Robins et al. 2007a, b; Robins 2009). The reasons that we employ the ERGM in our citation network analysis are twofold: on the one hand, as illustrated by Robins et al. (2007a), in ERGMs, not only individual-level attributes but also possible ties' surroundings could affect the occurrence of a relationship, which means both nodes' attributes and network structures (existing ties) are considered to "influence the generation of networks" (Zhang et al. 2018, p. 85); on the other hand, the ERGM shows the effects of the combination between local patterns of networks and social processes on the formation of global network patterns.

To maintain consistency, we use the notation shown in Zhang et al. (2018). Mathematically, the probability of observing the current network (w) is defined as:

$$P(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x})=\left(\frac{1}{\kappa}\right)\exp\left\{\begin{array}{l}\theta\sum_{i,j}y_{ij}+\left(m_{p_s}(\mathbf{y},\mathbf{x})+h_{p_s}(\mathbf{y},\mathbf{x})\right)\\+\left(m_{p\_f}(\mathbf{y},\mathbf{x})+h_{p\_f}(\mathbf{y},\mathbf{x})\right)+\left(m_{p_n}(\mathbf{y},\mathbf{x})+h_{p_n}(\mathbf{y},\mathbf{x})\right)\\+(m_c(\mathbf{y},\mathbf{x})+h_c(\mathbf{y},\mathbf{x}))+(m_t(\mathbf{y},\mathbf{x})+h_t(\mathbf{y},\mathbf{x}))\\+\left(m_g(\mathbf{y},\mathbf{x})+h_g(\mathbf{y},\mathbf{x})\right)\end{array}\right\} \tag{1}$$

where $\mathbf{Y}$ is a random citation network, $\mathbf{X}$ the covariates, $\mathbf{y}$ the observed citation network, and $\mathbf{x}$ the observed covariate; $\theta\sum_{i,j}y_{ij}$ the effects of the citation network's density; $m_{p\_s}(\mathbf{y},\mathbf{x})+h_{p\_s}(\mathbf{y},\mathbf{x})$ the main and homophily effects of authors' single-authored publication number;$m_{p\_f}(\mathbf{y},\mathbf{x})+h_{p\_f}(\mathbf{y},\mathbf{x})$ the main and homophily effects of authors' collaborative first-authored publication number; $m_{p\_n}(\mathbf{y},\mathbf{x})+h_{p\_n}(\mathbf{y},\mathbf{x})$ the main and homophily effects of authors' collaborative non-first-authored publication number; $m_c(\mathbf{y},\mathbf{x})+h_c(\mathbf{y},\mathbf{x})$ the main and homophily effects of authors' citation number; and $(m_t(\mathbf{y},\mathbf{x})+h_t(\mathbf{y},\mathbf{x}))$ the main and homophily effects of authors' top research interest;$m_g(\mathbf{y},\mathbf{x})+h_g(\mathbf{y},\mathbf{x})$ the main and homophily effects of authors' gender; κ the normalizing factor ensuring the probabilities sum to one (Robins et al. 2007a, b); and y_{ij} the variable indicating whether there is a tie between i and j ($y_{ij}=1$ or $y_{ij}=0$).

Table 1 Statistic descriptions for publication count and citation count in the citation network

Indicator	Max.	Min.	Avg.
Number of publications	36	6	≈ 9
Number of citations	3557	0	≈ 168

Results and discussion

The descriptive statistics for the selected 633 authors in the citation network are shown in Table 1, where the maximum number of publications a scholar in the network has written is 36, while the minimum number is 6. With the respect for the number of citations, the maximum and minimum numbers are 3557 and 0, respectively, which shows an extreme difference. The average number of publications and citations are approximately 9 and 168, respectively.

As for the topics we extracted with ACT model, we manually labeled them by carefully researching the authors as well as their publications, after which these topics were recognized as: (1) Database, (2) Medical Information Retrieval (IR), (3) IR Theory, (4) IR Systems, and (5) Image-based IR. The ERGM result is shown in Table 2, where the main effects and homophily effects of nodes' attributes as well as network structure are detailed separately.

Table 2 shows that the number of single-authored, collaborative and first-authored, and collaborative but not first-authored publications all have significant main effects on the formation of citation networks. Specifically, if we keep all other variables unchanged, the

Table 2 ERGM results for modeling the citation networks among the most productive authors

Variables	Est.	Sig.
Main effects		
No. of single-authored publications	0.08	***
No. of first-authored publications	0.06	***
No. of non-first-authored publications	0.03	**
No. of citations	0.18	***
Most-used Topic 2 (Medical IR)	0.32	
Most-used Topic 3 (IR Theory)	0.05	***
Most-used Topic 4 (IR Systems)	0.01	*
Most-used Topic 5 (Image-based IR)	0.08	
Gender (female)	0.06	
Homophily effects		
Single-authored publication no. difference	− 0.19	*
First-authored publication no. difference	− 0.02	
Non-first-authored publication no. difference	− 0.09	
Citation no. difference	− 1.03	***
Same most used topic	− 2.02	***
Same gender	− 0.09	*
Network structure		
Edges	− 2.53	***

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

probabilities of obtaining a new citation for one scholar having one more unit of single-authored, collaborative and first-authored, as well as collaborative but non-first-authored publications are 1.08 ($e^{0.08}$), 1.06 ($e^{0.06}$), and 1.03 ($e^{0.03}$) times, respectively, compared with other authors ($p < 0.001$ for the first two variables, and $p < 0.01$ for the third variable). These strong main effects illustrate the potential proportional relationship between the number of publications and the obtaining of citations based on the estimate results from the ERGM. Additionally, the estimated value for number of single-authored publications is the largest among the three, indicating that solo-authored publications might be able to facilitate authors being cited more by others. Based on this result, writing single author publications, perhaps, are more effective at gaining citations for their authors than collaborative ones. In the era of "collaborative imperative" (Bozeman and Boardman 2014), people tend to divide their labors into several parts to collaboratively publish an article (Boyer-Kassem and Imbert 2015), and the ratio of single-authored papers are decreasing (Wuchty et al. 2007). However, to recognize the value of single-authored paper could be of importance because it is more likely to show the versatility of the solo author since he/she finishes most of the tasks (e.g., various tasks such as looking for related studies, writing papers, doing empirical studies, and collecting data, etc.) related to the paper by himself/herself. Another thing worth noticing is the negative homophily effects ($p < 0.05$) of number of single-authored publications on the formation of citation networks. Particularly, there is only about 82.7% ($e^{-0.19}$) chance that researchers will select another researcher who has published similar number of single-authored publications with him/her to be cited, compared with selecting another researcher whose single-authored publications are quite different. The counter-homophily effects of citation numbers reveal the potential reluctance of citing authors with similar solo-authored articles, which is probably indicative of the states of scientific careers or academic titles to some extent. Although we are not sure whether researchers really consider these factors when they cite papers, this finding reveals some potential subconscious influences of scholars and it calls for future research from the perspectives of psychology and cognitive science from both qualitative and quantitative ways.

From Table 2 we can see that the number of citations has very significant main ($p < 0.001$) and homophily effects ($p < 0.001$) on the formation of citation network structures. Specifically, as for the main effect, when adding one unit of citation number, an author will be 19.7% ($e^{0.18} - 1$) more likely to be cited once again than others, if we keep all other variables fixed. This finding confirms many previous findings such as preferential attachment (Newman 2001; Barabási et al. 2002; Jeong et al. 2003; Wang et al. 2008, 2009; Milojević 2010; Wang et al. 2013), "Matthew effect" (Merton 1968), or the "cumulative advantage" (de Solla Price 1976), as it reinforces the notion that the more number of citations an author has received, the more likely they have to receive new citations. Meanwhile, the significant negative homophily effect estimates illustrate that there is only a 35.7% ($e^{-1.03}$) chance that authors will cite another author having received similar number of citations, compared with selecting an author whose number of citations is very different. That is to say, reputation of authors or articles makes a difference to citing authors in selecting references, which is commonly expressed "success breeds success" (Cozzens 1985). Authors with low number of citations could have more possibilities to cite authors who have high number of citations (larger difference in terms of citation number between citing and cited authors) than to cite those low (smaller difference in terms of citation number between citing and cited authors). This significant counter-homophily effect confirms several previous findings in which high-impact authors, often measured by the high number of citations, have tendencies to be cited more than others (e.g., Van Raan et al. 2003).

In terms of most frequently used research topic, a typical categorical variable instead of count variable, we simply compare the estimated probabilities of gaining citation for the authors whose most frequent research topics are Medical IR (Topic 2), IR Theory (Topic 3), IR Systems (Topic 4), and Image-based IR (Topic 5) with authors who mostly prefer the research topic of Database (Topic 1). The results show that the possibilities that researchers in Topics 3 or 4 are cited are increased by 5.1% ($e^{0.05} - 1$) ($p < 0.001$) and 1.0% ($e^{0.01} - 1$) ($p < 0.05$), which means that the authors in these fields are generally likely to receive more citations; the effect of the former (i.e., for Topic 3) is relatively obvious. Such findings could be acceptable, for IR theory articles could receive more citations because they could be used in many general IR studies as a foundation. The homophily effect of research topic on the formation of citation networks is, on the other hand, significant ($p < 0.001$). Specifically, the probability of citing others sharing the same frequently used research topic in the field of information retrieval is 7.54 ($e^{2.02}$) times if we keep all other features fixed, which is a very natural result that could also be confirmed by several previous studies (e.g., Cases and Higgins 2000; Chubin and Moitra 1975). For instance, the study provided by Chubin and Moitra (1975) maintained seven reasons for citations, where describing relevant work and providing historical background were attached much significance.

As for gender, we fail to detect any significant main effects but find a weak but positive homophily effect ($p < 0.05$) on the citation network formation. The possibility of citing same-gender authors is 1.09 times compared with that of citing cross-gender ones. Nevertheless, such finding could result from the limited number of females (only $\approx$ 12%) among the most productive authors selected in the IR field, which makes limited number of cross-gender citing counts. Also, it could reflect the lack of highly cited females in the IR field itself. Additionally, the imbalance of gender demonstrated a potential citing reason, for male are more common to be cited than female (Cole and Singer 1991; Baldi 1998; Stack 2004). Moreover, Table 2 also shows an obvious negative estimate for "edge" ($p < 0.001$), which indicates the density of our constructed citation network is much smaller than other randomly generated networks with similar features—our constructed network is actually sparse.

Conclusions

As a supplement of previous qualitative research, this paper aims to quantitatively explore several potential citing reasons by analyzing factors including the number of publications (detailed by the number of single-authored, collaborative and first-authored, collaborative but non-first-authored publications), number of citations, most frequently used research topic, and gender. By employing the Exponential Random Graph Model (ERGM), we successfully examined the main and homophily effects simultaneously in the same model and considered these effects from the perspective of cited authors quantitatively. The results of our study show that the number of single-authored, collaborative and first-authored, and collaborative but non-first-authored publications all have significant main effects on the construction of citation networks, while authors tend to cite those with different number of single-authored publications from them. Furthermore, the number of citations has a very significant main and homophily effects on the formation of citation networks. In terms of the most frequently used research topic, publications containing basic theories about information retrieval are more likely to be referred to in the experiment data set. Moreover, the more similar their research topics are, the more probabilities to be cited. As for gender, we fail to detect any main effect but find a weak but positive homophily effect on the citation network formation.

However, there are several limitations in this article. Firstly, we simply consider the field of information retrieval instead of more various disciplines. Also, from the perspective of the dataset, the incomplete records in WoS database might negatively affect the performance of author name disambiguation and thus influence on the final results, a point also noted by Zhang et al. (2018). The calculation of the publication count is also limited based on the dataset. An author who has published many papers in other fields other than just information retrieval might be underestimated. Although the percolation procedure of selecting the top 500 productive authors can to some extent solve this, we still believe that this is one of the limitations of the current study. Future researchers might want to use different and more diverse datasets, such as the whole WoS or Microsoft Academic Graph (MAG) datasets (Sinha et al. 2015) to improve this.

Methodologically, we only considered a limited number of attributes concerning cited authors. In the future, we can apply several other indicators, such as h-index (Hirsch 2005), the nationality, and native language of authors by using Ethnea (Torvik and Agarwal 2016), a platform that provides the latter two pieces of information by inputting the authors' full name. Meanwhile, some network structures besides network density could be involved, such as transitivity (Newman and Park 2003) and k-star (Aghagolzadeh et al. 2012) of citation network, which will definitely expand our perspectives of regarding citation network as well as citing behaviors in future studies.

Acknowledgements The authors would like to thank the anonymous reviewers and Dr. Wolfgang Glänzel for their kind help on improving the quality of this paper.

References

Aghagolzadeh, M., Barjasteh, I., & Radha, H. (2012). Transitivity matrix of social network graphs. In *Proceedings of the 2012 IEEE statistical signal processing workshop* (pp. 145–148), August 5–8, 2012, Ann Arbor, MI, USA.

Baldi, S. (1998). Normative versus social constructivist processes in the allocation of citations: A network-analytic model. *American Sociological Review, 63*(6), 829–846.

Barabási, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications, 311*(3), 590–614.

Bornmann, L., & Daniel, H. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45–80.

Boyer-Kassem, T., & Imbert, C. (2015). Scientific collaboration: Do two heads need to be more than twice better than one? *Philosophy of Science, 82*(4), 667–688.

Bozeman, B., & Boardman, C. (2014). *Research collaboration and team science: A state-of-the-art review and agenda*. New York, PA: Springer.

Brooks, T. A. (1986). Evidence of complex citer motivations. *Journal of the American Society for Information Science, 37*(1), 34–36.

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science, 40*(4), 284–290.

Carpenter, M. P., & Narin, F. (1981). The adequacy of the Science Citation Index (SCI) as an indicator of international scientific activity. *Journal of the American Society for Information Science, 32*(6), 430–439.

Cases, D. O., & Higgins, G. M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the Association for Information Science and Technology, 51*(7), 635–645.

Chen, C. (2012). Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology, 63*(3), 431–449.

Chubin, D., & Moitra, S. (1975). Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science, 5*(4), 423–441.

Cole, S., & Singer, B. (1991). A theory of limited differences. In H. Zuckerman, J. R. Cole, & J. T. Bruer (Eds.), *The outer circle: Women in the scientific community*. London: W.W. Norton and Company.

Cozzens, S. E. (1985). Comparing the sciences: Citation context analysis of papers from neuropharmacology and the sociology of science. *Social Studies of Science, 15*(1), 127–153.

Cronin, B. (1984). *The citation process: The role and significance of citations in scientific communication*. London: Taylor Graham.

de Solla Price, D. J. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*(5), 292–306.

Ding, Y., & Cronin, B. (2011). Popular and/or prestigious? Measures of scholarly esteem. *Information Processing and Management, 47*(1), 80–96.

Duncan, E. (1981). Qualified citation indexing: Its relevance to educational technology. In E. Duncan and R. McAleese (Eds.), Information retrieval in educational technology. *Proceedings of the first symposium on information retrieval in educational technology,* April 1, 1981, Aberdeen (pp. 70–79). Aberdeen, Scotland: University of Aberdeen.

Erikon, M. G., & Erlandson, P. (2014). A taxonomy of motives to cite. *Social Studies of Science, 44*(4), 625–637.

Garfield, E. (1955). Citation Indexes for science: A new dimension in documentation through association of ideas. *Science, 122*(3159), 108–111.

Garfield, E. (1964). Can citation indexing be automated? *In Proceedings symposium of the statistical association methods for mechanized documentation* (pp. 2–4), March 17–19, 1964, Washington D.C.

Garfield, E. (1965). Can citation indexing be automated. In *Statistical association methods for mechanized documentation, symposium proceedings* (Vol. 269, pp. 189–192). Washington, DC: National Bureau of Standards, Miscellaneous Publication 269

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science, 178*(4060), 471–479.

Garfield, E. (1998). Random thoughts on citationology, its theory and practice: Comments on theories of citation. *Scientometrics, 43*(1), 67–76.

Gilbert, G. N. (1977). Referencing as persuasion. *Social Studies of Science, 7*(1), 113–122.

Gross, P. L. K. (1927). College libraries and chemical education. *Science, 66*(1713), 385–389.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., & Morris, M. (2003). *Statnet: Software tools for the statistical modeling of network data*. Retrieved April 1, 2017, from http://statnetproject.org.

Hendley, M. (2012). Citation behavior of undergraduate students: A study of history, political science, and sociology papers. *Behavioral and Social Sciences Librarian, 31*(2), 96–111.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceeding of the National Academy Science of United States of America, 102*(46), 16569–16572.

Jeong, H., Néda, Z., & Barabási, A. L. (2003). Measuring preferential attachment in evolving networks. *Europhysics Letters, 61*(4), 567–572.

Kapseon, K. (2004). The motivation for citing specific references by social scientists in Korea: The phenomenon of co-existing references. *Scientometrics, 59*(1), 79–93.

Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature, 411*(6837), 521.

Li, Z., Peng, Q. K., & Liu, C. (2016). Two citation-based indicators to measure latent referential value of papers. *Scientometrics, 108*(3), 1299–1313.

Lipetz, B. A. (1965). Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. *Journal of the Association for Information Science and Technology, 16*(2), 81–90.

Liu, M. (1993). Study of citing motivation of Chinese scientists. *Journal of Information Science, 19*(1), 13–23.

Marx, W., & Bornmann, L. (2015). On the causes of subject-specific citation rates in Web of Science. *Scientometrics, 102*(2), 1823–1827.

McDowell, J. M., & Smith, J. K. (1992). The effect of gender-sorting on propensity to coauthor: Implications for academic promotion. *Economic Inquiry, 30*(1), 68–82.

Merton, R. K. (1957). Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review, 22*(6), 635–659.

Merton, R. K. (1968). The Matthew effect in science. *Science, 159*(3810), 56–63.

Milojević, S. (2010). Modes of collaboration in modern science: Beyond power laws and preferential attachment. *Journal of the American Society for Information Science and Technology, 61*(7), 1410–1423.

Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical Review E, 64*(2), 025102.

Newman, M. E., & Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E, 68*(3), 036122.

Prabha, H. G. (1983). Some aspects of citation behavior: A pilot study in business administration. *Journal of the American Society for Information Science, 34*(3), 202–206.

Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007a). An introduction to exponential random graph (p*) models for social networks. *Social Networks, 29*(2), 173–191.

Robins, G., Pattison, P., & Wang, P. (2009). Closure, connectivity and degree distributions: Exponential random graph (p*) models for directed social networks. *Social Networks, 31*(2), 105–117.

Robins, G., Snijders, T., Wang, P., Handcock, M., & Pattison, P. (2007b). Recent developments in exponential random graph (p*) models for social networks. *Social Networks, 29*(2), 192–215.

Sandstrom, U., Wadskog, D., & Karlsson, S. (2005), Research institutes and universities: does collaboration pay? In P. Ingwersen and B. Larsen (Eds.), *Proceedings of the tenth international conference of the international society for scientometrics and informetrics,* Karolinska University Press, Stockholm.

Silverman, R. J. (1985). Higher education as a maturing field? Evidence from referencing practices. *Research in Higher Education, 23*(2), 150–183.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B. J. P., & Wang, K. (2015). An overview of Microsoft Academic Service (MAS) and applications. In *Proceedings of the 24th international conference on World Wide Web* (pp. 243–246), May 18–22, 2015, Florence.

Soper, M. E. (1976). Characteristics and use of personal collections. *Library Quarterly, 46*(4), 397–415.

Stack, S. (2004). Gender, children and research productivity. *Research in Higher Education, 45*(8), 891–920.

Tang, J., Jin, R., & Zhang, J. (2008). A topic modeling approach and its integration into the random walk framework for academic search. *In Proceeding of the eighth IEEE international conference on data mining* (pp. 1055–1060), December 15–19, 2008, Pisa.

Torvik, V. I., & Agarwal, S. (2016). Ethnea: An instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database. In *Proceedings of the international symposium on science of science*, March 22–23, 2016, Library of Congress, Washington D.C.

Van Raan, A. F. J., Visser, M. S., Van Leeuwen, T. N., & Van Wijk, E. (2003). Bibliometric analysis of psychotherapy research: Performance assessment and position in the journal landscape. *Psychotherapy Research, 13*(4), 511–528.

Wang, M., Li, S., & Chen, G. (2017). Detecting latent referential articles based on their vitality performance in the latest 2 years. *Scientometrics, 112*(3), 1557–1571.

Wang, D., Song, C., & Barabási, A. L. (2013). Quantifying long-term scientific impact. *Science, 342*(6154), 127–132.

Wang, M., Yu, G., & Yu, D. (2008). Measuring the preferential attachment mechanism in citation networks. *Physica A: Statistical Mechanics and its Application, 387*(18), 4692–4698.

Wang, M., Yu, G., & Yu, D. (2009). Effect of the age of papers on the preferential attachment in citation networks. *Physica A: Statistical Mechanics and its Applications, 388*(19), 4273–4276.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p. *Psychometrika, 61*(3), 401–425.

Weinstock, M. (1971). Citation indexes. *Encyclopedia of Library and Information Science, 5,* 16–40.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5827), 1036–1039.

Yin, Y., & Wang, D. (2017). The time dimension of science: Connecting the past to the future. *Journal of Informetrics, 11*(2), 608–621.

Yu, Q., Long, C., Lv, Y., Shao, H., & He, P. (2014). Predicting co-author relationship in medical coauthorship networks. *PLoS ONE, 9*(7), e101214.

Zhang, C., Bu, Y., Ding, Y., & Xu, J. (2018). Understanding scientific collaboration: Homophily, transitivity, and preferential attachment. *Journal of the Association for Information Science and Technology, 69*(1), 72–86.

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in Information Science 1996–2005: Introducing author bibliographic coupling analysis. *Journal of the American Society for Information Science and Technology, 59*(13), 2070–2086.

Zhao, D., & Strotmann, A. (2014). The knowledge base and research front of Information science 2006-2010: An author co-citation and bibliographic coupling analysis. *Journal of the Association for Information Science and Technology, 65*(5), 996–1006.