Understanding Scientific Collaboration:

Homophily, Transitivity, and Preferential Attachment

Chenwei Zhang and Yi Bu

Indiana University Bloomington

Ying Ding

Indiana University Bloomington

Tongji University

Wuhang University

Jian Xu

Sun Yat-sen University

Author Note

Chenwei Zhang, Department of Information and Library Science, Indiana University Bloomington, Bloomington, IN, U.S.A.; Yi Bu, School of Informatics and Computing, Indiana University Bloomington, Bloomington, IN, U.S.A.; Ying Ding, School of Informatics and Computing, Indiana University Bloomington, Bloomington, IN, U.S.A., University Library, Tongji University, Shanghai, China, School of Information Management, Wuhan University, Wuhan, HuBei, China; Jian Xu, School of Information Management, Sun Yat-sen University, Guangzhou, Guangdong, China.

Correspondence concerning this article should be addressed to Ying Ding, School of Informatics and Computing, Indiana University, Bloomington, IN 47408. E-mail: dingying@indiana.edu

Abstract

Scientific collaboration is essential in solving problems and breeding innovation. Coauthor network analysis has been utilized to study scholars' collaborations for a long time, but these studies have not simultaneously taken different collaboration features into consideration. In this paper, we present a systematic approach to analyze the differences in possibilities that two authors will cooperate as seen from the effects of homophily, transitivity, and preferential attachment. Exponential random graph models (ERGMs) are applied in this research. We find that different types of publications one author has written play diverse roles in his/her collaborations. An author's tendency to form new collaborations with his/her coauthors' collaborators is strong, where the more coauthors one author had before, the more new collaborators he/she will attract. We demonstrate that considering the authors' attributes and homophily effects as well as the transitivity and preferential attachment effects of the coauthorship network in which they are embedded helps us gain a comprehensive understanding of scientific collaboration.

Keywords: scientific collaboration, coauthorship network, homophily, transitivity, preferential attachment, exponential random graph models

Understanding Scientific Collaboration: Homophily, Transitivity, and Preferential Attachment

Scientific collaboration makes the impossible possible. The Human Genome Project (HGP) is the world's largest collaborative biological project which has over twenty universities and research centers located in six countries. The outcome of this project has provided breakthroughs in fields ranging from molecular medicine to human evolution. Collaboration almost becomes mandatory in many fields, where success of research heavily depends on team work. More than 90% of publications in science, technology, and engineering are found to be collaborative (Bozeman & Boardman, 2014). Collaboration also breeds innovation. For example, in 2014, the international collaboration of scientists from 20 countries has unveiled the myth of the genetic basis of schizophrenia that affects nearly 24 million people globally (Flint & Munafò, 2014). Understanding the mechanisms and processes of scientific collaboration is therefore critical, especially when determining how to develop breakthrough innovations.

Theoretically, from the network science perspective, scientific collaborations among scholars form a social network (Newman & Park, 2003). There exist several fundamental mechanisms by which the network forms and evolves. Homophily is a fundamental effect in social networks to describe how people have a tendency to make connections in the networks with those who have similarities to themselves (McPherson, Smith-Lovin, & Cook, 2001). Such a mechanism will force scholars to form more homogeneous collaboration with respect to the authors' characteristics. Homophily has been observed among a broad range of collaborations (Boschini & Sjögren, 2007; Freeman & Huang, 2014; Sie, Drachsler, Bitter-Rijpkema, & Sloep, 2012).

Transitivity is another common phenomenon. It means that there is a high probability of two nodes being connected if they are connected to one (or more) common nodes. There is usually a high degree of transitivity in social networks (Newman, 2001a). Such a mechanism will make the collaboration to be path dependent. The scholars will follow their connections to find collaborators by linking to their coauthors' coauthors. Transitivity has been widely examined in the area of scholarly collaboration (Newman, 2001a; Franceschet, 2011; Schilling and Phelps, 2007).

During the evolution of the collaboration network, preferential attachment also plays an important role. As a key feature of real network (Barabási and Albert, 1999), it refers to that the more existing ties one node has, the more new connections it is likely to accumulate. It is related to the theory of cumulative advantage in science, known as the "Matthew effect," (Merton, 1968; de Solla Price, 1976). It infers that the ability to gain collaborators may increase with the scholars' centralities in the network. The preferential attachment process generates a "long-tailed" distribution following a Pareto distribution or power law in its tail, a phenomenon that has been extensively demonstrated in collaboration networks (Newman, 2001a; Barabási et al., 2002; Jeong, Néda, and Barabási, 2003).

From the perspective of collaboration theories, both the mechanisms of homophily and transitivity are also related to the important factors deciding collaboration—the search cost and the communication cost (Boudreau et al., 2014; Kraut, Egido, & Galegher, 1988). Homophily encourages people sharing similar backgrounds to work together, thus they tend to have less barriers in communication. Transitivity provides the scholars a direction to find their potential collaborators, rather than by random selection, which may cost more in searching and matching. In addition, the preferential attachment also reflects one of the key motivations (Katz, 1994) for a researcher to collaborate—by cooperating with those famous scholars, he/she is more likely to be more productive, visible and recognized.

Most scholars need to make daily decisions about selecting potential collaborators or accepting collaboration invitations from others. Previous studies have shown the mechanisms of homophily, transitivity, and preferential attachment (e.g., Newman, 2001a; Barabási et al., 2002; Moody, 2004; Boschini & Sjögren, 2007; Franceschet, 2011; Freeman & Huang, 2014) could all influence the decisions of scientific collaboration. Yet these features were generally introduced as static indicators, and were examined in isolation. In network formation, the interdependent nature of different features has been confirmed. The creation of one connection may affect others, all of which needed to be "considered jointly for proper inference" (Goodreau, Kitts, & Morris, 2009, p. 104), and where one observation may be the result of different effects. In reality, scientific collaboration is affected by various factors wherein their influences are simultaneous. If only examining separately, we could not conclude how strong every effect contributes to the generation of scientific collaboration simultaneously in an integrated environment. Until recently, studies have not clearly addressed the question of how strongly these features affect two scholars' scientific collaboration. Given the various explanations, it can be useful if we could understand how these determinants work together to influence scholars' collaboration decisions; we want to know how to set the different criteria to select collaborators when encountering various situations in real life. This study presents a systematic approach to analyze the differences in possibilities that two authors will cooperate based on the effects of homophily, transitivity, and preferential attachment simultaneously. The homophily is examined based on the input of several authors' attributes; while the transitivity and preferential attachment are investigated by the whole network structure thus they do not depend on the input of exogenous data. We are able to find out the real effects of each factor, when all other factors exist, rather than overestimating a certain factor by ignoring all other factors.

An exponential random graph model (ERGM) is employed to model the network formation (Wasserman & Pattison, 1996; Robins, Pattison, Kalish, & Lusher, 2007a; Robins, Snijders, Wang, Handcock, & Pattison, 2007b; Robins, Pattison, & Wang, 2009) with the simultaneous effects of both individual authors' attributes and network structures (Goodreau et al., 2009). This model measures the generation of authors' collaboration relationships as a stochastic process and incorporates both covariate effects of authors' attributes and social network structure features to understand research collaboration, rather than examining each feature in isolation and static. It reflects the formation of the realistic collaboration network and helps us distinguish similar patterns observed in the collaboration network which are caused by different features. Meanwhile, this model allows us to calculate the possibilities that two authors might collaborate resulted from the effects various features (i.e., homophily, transitivity, and preferential attachment). A detailed illustration of the ERGM method is provided in the Appendix (available upon request).

In this present study, we address the following three questions. First, in scientific collaboration, whether the effects of authors' attributes and the structure of the collaboration network itself simultaneously contribute to the formation and evolution of the collaboration networks? Second, what are the roles that the homophily based on the authors' attributes, the transitivity and the preferential attachment play in the process of network formation? Third, how do homophily, transitivity, and preferential attachment help us better understand scientific collaboration? This paper is outlined as follows: Section 1 introduces scientific collaboration research; Section 2 provides the literature review of scientific collaboration and; Section 3 explains the data collection and method used in this paper and proposes our hypotheses; Section

4 discusses the results; and Section 5 draws the conclusion and points out some future research directions.

Related Works

Collaboration and Authors' Attributes: Productivity, Impact, Research Interests, and Gender

The relation between authors' productivities and their collaboration has been demonstrated in many fields. For example, de Solla Price and Beaver (1966), when investigating the collaboration of memos (informal publications, mostly are preprints of articles) between members of an information exchange group in health related domains, found that the more prolific one author is, the more collaboration he was involved in. Similar results were also found by Pravdić and Oluić-Vuković (1986). Based on the study of the curricula vitae and surveys of 443 scholars associated with the National Science Foundation or Department of Energy, Lee and Bozeman (2005) showed a strong correlation between one scientist's publishing productivity, and the number of collaborators he had. Even in humanity disciplines, such as musicology, Pao (1982) found the two most productive musicologists were also the most collaborative. But in most studies, such relations were simply studied by correlation, where the analysis was descriptive, suggesting that this research would benefit from more in-depth examination. In addition, most research was conducted in the direction that scientific collaboration leads to productivity. In this work, we argue that this influence works both ways, in that the more productive one scholar is, the more other researchers may tend to collaborate with him/her. We thus investigate the effect of authors' productivity on their collaboration.

Most efforts that investigated the relation between scholarly collaboration and impact were found to be at the article level, such as in examining how collaboration contributes to the increase of citations of a work. Leimu and Koricheva (2005) analyzed the citation rates of works resulting from different types of collaborations in the field of ecology. They found the influence of collaboration on the impact of the resulting work is not always positive or even minor in general. Thurman and Birkinshaw (2006) found that the number of citations was significantly associated with the number of coauthors in six leading journals in medicine. When Hsu and Huang (2010) explored the correlations between the number of citations and the number of coauthors in eight scientific journals, they found that “predicting the citation number from the coauthor number can be more reliable than predicting the coauthor number from the citation number” (p. 317). Focusing on authors rather than the articles here, we explore whether the authors’ impact (the number of citation he/she receives) makes any difference on the number of collaborations he/she has, and include the effects of authors’ citation numbers on their collaborations.

Studies that have demonstrated the relation between collaboration and authors’ research interests include that of Kraut, Egido, and Galegher (1988), who pointed out that sharing research similarities encourages scientific collaboration, and Ding (2011), who found productive authors in the information retrieval field have a tendency to coauthor with those who share similar research interests with them. A few studies investigated authors’ collaboration patterns at the topic level within one broad domain. For example, Huang, Zhuang, Li, and Giles (2008) generated coauthorship networks in six topics from CiteSeer data and contrasted the collaboration characteristics. In this paper, we examine the effect of authors’ research interests on their collaborations and further explore these collaborations in each topical sub-graph.

By examining a cohort sample of Ph.D. economists, McDowell and Smith (1992) found that researchers tend to collaborate with those of the same sex. When modeling the coauthorship patterns during 1991-2002 in three top economics journals, Boschini and Sjögren (2007) found that females tend to collaborate with the same gender authors. In this paper, we also explore the role gender plays on researchers to form collaboration.

Homophily in Scientific Collaboration

A few scholars confirmed the effects of homophily in coauthorship patterns. Boschini and Sjögren (2007), in investigating coauthorship patterns in articles published during 1991-2002 in three top economics journals, found that women were two times as likely as men to collaborate with women; and the female-male gap in the propensity to collaborate with a female author increases with the presence of women. Sie et al. (2012) noted the importance of authors' similarities when forming collaborations, and thus adopted the similarities between authors' keywords as a rule for suggesting future co-authors for scientific paper writing. The evaluation showed this similarity-based method to be feasible. Similar studies include that of Freeman and Huang (2014) for homophily on authors' ethnicity and Boschini & Sjögren (2007) for authors' sex. In this paper, we analyze the homophily effect based on the collaboration graph, where all the coauthors are examined to show how the homophily mechanism influences the evolution of collaboration network.

Transitivity in Scientific Collaboration

The collaboration network is a type of social network where transitivity has been widely investigated. Newman analyzed the transitivity of coauthorship in a few domains such as biology, physics, and mathematics and computer science (Newman, 2001a, 2001b, 2001c, 2004). He used the clustering coefficient to quantify the networks' transitivity, and found that "the probability of a pair of scientists collaborating increases with the number of other collaborators they have in common" (Newman, 2001a, p. 1). From an analysis of collaboration in the field of computer science since 1936, including both journal publications and conference articles, Franceschet (2011) found the chance that two researchers who share common collaborators in a publication was quite high. He also found the transitivity of journal publication coauthorship networks was even higher than that of conference proceeding collaboration networks.

As transitivity is a small-scale characteristic of the social networks (Newman & Park, 2003), the transitivity index has been used as "a global metric quantifying the tendency of this small-scale attribute over the entire graph; it is proportional to ratio of the number of triangles over the total number of connected triples" (Aghagolzadeh, Barjasteh, & Radha, 2012, p. 145). How a network's transitivity affects the formation of its ties is important, but the transitivity index in these studies was only a static index and could not reveal the variations of this characteristic in the network (Aghagolzadeh et al., 2012). Instead of showing a static index of transitivity, this paper measures how this transitive structure precisely influences the emergence of new scholarly collaborations.

Preferential Attachment in Scientific Collaboration

Preferential attachment has been widely known to influence the generation of new scholarly collaborations. Newman (2001a) analyzed the preferential attachment in coauthor networks in physics and biology and found the number of new collaborations one author gained each year increased with the number of his past collaborators. Barabási et al. (2002) demonstrated the presence of preferential attachment in two collaboration networks in mathematics and neuroscience for an eight-year period and found the emergence of a new publication was more likely to occur among those who already had a large number of coauthors. Jeong et al. (2003) in measuring the preferential attachment effect found the attachment rate is sublinear in the coauthorship network of neuroscience. Milojević (2010) found that authors in nanoscience with more than twenty collaborators benefit from preferential attachment when forming new coauthorships. In this paper, we examine the precise effect of the preferential attachment process in scientific collaboration.

Methodology and Hypotheses

Data

Papers and their corresponding citations for this paper are harvested from Web of Science (WoS) in the time range of 1956-2014. Information retrieval is selected as the testing field. Information retrieval is a subdomain in Computer Science and a transdisciplinary field. According to Franceschet and Costantini (2010), the scholars in the field of computer science produce more valuable papers with moderate collaboration. The coauthorship in a computer science paper demonstrates one author has played a substantial role in this publication (Solomon, 2009). Unlike disciplines that usually have a large list of coauthors, such as biomedicine and high-energy physics (Cronin, 2001), every co-author in one publication found in Information retrieval has a significant level of involvement in the collaboration. We refer to Ding (2011) for a list of query terms. The dataset contains 59,162 authors who published 20,359 papers, in which there are 558,498 references. To disambiguate author names, a simple two-step matching procedure based on author name and affiliation (Yu et al., 2014) is employed. After applying their method, we identify 44,770 distinct authors in the dataset.

According to the literature, we already know that the number of publications and the number of citations one author has are associated with their levels of collaboration. Thus we incorporate their effects in this study. In addition, we investigate the effects of collaboration on an author's different types of publications—single-authored, collaborating and serving as the first author, and collaborating but as a non-first author. We thus collected these three variables separately. Meanwhile, as we examine the effects of the authors' research similarities on their collaborations, the top research interest of each author is included. We are also interested in how the authors' gender (McDowell & Smith, 1992) plays a role in their collaboration, thus we collected the authors' gender information. In total, we collected the following six attributes for each author:

1. The number of single-authored papers one author published (count variable);
2. The number of collaborating-first-authored papers one author published (count variable);
3. The number of collaborating-non-first-authored papers one author published (count variable);
4. The number of citations one author's all publications received (count variable); and
5. The most frequently used topic (categorical variable).
6. The gender information (categorical variable).

Methods

Coauthorship networks. We first rank all the authors by the number of papers each author published. Initially we wanted to select the top 500 productive authors. Since the authors numbered from the 447th to the 633rd all have published six papers, we include all these 633 most productive authors in the dataset. Each author represents one node in the network. If two authors have collaborated in one paper, a tie is added between them. We do not consider the frequency of collaborations between two authors as the weight of their tie, so that the network is binary. We use the Author-Conference-Topic (ACT) model by Tang, Jin and Zhang (2008) to extract the authors' research topic distribution. We set the number of topics to extract as five and use the topic with the highest weight in each author's distribution as his/her core research interest. If there is more than one topic having the same highest weight, we randomly select one of them for the dataset.

Exponential random graph models. We apply ERGMs to model the coauthorship network and their attributes, where the probability of observing the current network (w) is:

$$\Pr(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}) =$$

$$\left(\frac{1}{\kappa}\right) \exp \left\{ \begin{array}{c} \theta \sum_{i,j} y_{ij} + \eta_1 C(\boldsymbol{y}) + \eta_2 U(\boldsymbol{y}) + \left(m_p(\mathbf{y},\mathbf{x}) + h_p(\mathbf{y},\mathbf{x})\right) + \\ \left(m_c(\mathbf{y},\mathbf{x}) + h_c(\mathbf{y},\mathbf{x})\right) + \left(m_t(\mathbf{y},\mathbf{x}) + h_t(\mathbf{y},\mathbf{x})\right) + \left(m_g(\mathbf{y},\mathbf{x}) + h_g(\mathbf{y},\mathbf{x})\right) \end{array} \right\} \qquad (1)$$

where **Y** is a random network, **X** the covariates, and **y** the observed network; $\theta \sum_{i,j} y_{ij}$ the effects of the network's density; $\eta_1 C(\boldsymbol{y})$ the effects of transitivity in the network's structures; $\eta_2 U(\boldsymbol{y})$ the effects of preferential attachment; $m_p(\mathbf{y},\mathbf{x}) + h_p(\mathbf{y},\mathbf{x})$ the main and homophily effects of authors' publication number; $m_c(\mathbf{y},\mathbf{x}) + h_c(\mathbf{y},\mathbf{x})$ the main and homophily effects of authors' citation number; $\left(m_t(\mathbf{y},\mathbf{x}) + h_t(\mathbf{y},\mathbf{x})\right)$ the main and homophily effects of authors' top research interest; and $\left(m_g(\mathbf{y},\mathbf{x}) + h_g(\mathbf{y},\mathbf{x})\right)$ the main and homophily effects of authors' gender.

Making a transformation of the general ERGM form in Equation (1), we obtain the following conditional logit model (Wasserman & Robins, 2005; Robins et al., 2007a):

$$\log \left[\frac{\Pr(Y_{ij} = 1 | \mathbf{y}_{\mathrm{ij}}^{C})}{\Pr(Y_{ij} = 0 | \mathbf{y}_{\mathrm{ij}}^{C})} \right] = \sum_{A(Y_{ij})} \eta_A \, d_A(\mathbf{y}) = \sum_{A(Y_{ij})} \eta_A \left[g_A\left(Y_{ij}^{+}, \mathbf{x}\right) - g_A\left(Y_{ij}^{-}, \mathbf{x}\right) \right] \qquad (2)$$

where the sum is over all configurations A that contain Y_{ij}; $d_A(\mathbf{y})$ is the change of network statistic; where it measures the difference between the network statistic when Y_{ij} is present ($g_A\left(Y_{ij}^{+}, \mathbf{x}\right)$) and when Y_{ij} does not exist ($g_A\left(Y_{ij}^{-}, \mathbf{x}\right)$); η_A is the corresponding parameter; and $\mathbf{y}_{\mathrm{ij}}^{C}$ is the rest of the observed network except the tie Y_{ij}.

From Equation (2), we understand that the logarithm of the ratio of the probability that a tie Y_{ij} is formed to the probability that Y_{ij} is not formed is equals to the changes of any covariate or local network structure when the tie Y_{ij} is flipped from 0 to 1. The coefficients in the ERGM are interpreted like this, which we call "log odds." For example, if the coefficient of one effect is β, we could say the possibility of creating a tie Y_{ij} is e^{β} times of the possibility of not creating such a tie, according to the changes brought by one unit of difference in this certain effect.

Hypotheses

Based on existing literature, we propose our hypotheses as below:

H1. Homophily effect plays an important role in the formation of scientific collaboration network.

Based on the existing literature reviewed above, we further specify this hypothesis into four different hypotheses:

H1a. Homophily effect measured by the authors' productivity influences the generation of collaboration ties.

H1b. Homophily effect measured by the authors' impact influences the generation of collaboration ties.

H1c. Homophily effect measured by the authors' research topics influences the generation of collaboration ties.

H1d. Homophily effect measured by the authors' gender influences the generation of collaboration ties.

H2. Transitivity effect plays an important role in the formation of scientific collaboration network.

H3. The effect of preferential attachment plays an important role in the formation of scientific collaboration network.

Results and Discussion

Overview

We first examine the number of authors' publications and the number of citations they received. In the network, the maximum number of publications one author has is 36, while the minimum is 6. On average, each author has written about 9 papers. One author has written at most 16 single-authored papers; 22 collaborating-first-authored papers; and 28 collaborating-non-first-authored papers. Some authors did not write any articles individually, or did not coauthor with others at all. The highest number of citations one author received is 3,557 and the lowest is 0. The average number of citations is more than 168. We manually label the five topics extracted from the author-topic-modeling as: Database (Topic 1), Medical Information Retrieval (Topic 2), Information Retrieval Theory (Topic 3), Information Retrieval Systems (Topic 4), and Image-based Information Retrieval (Topic 5).

ERGM Results on the Whole Collaboration Network

It is worth noting that in this study, the weights of authors' collaboration are ignored. We care about whether one author collaborates with different other authors, but we do not care about the strength or degree of collaboration, that is, how one researcher coauthors with another one repeatedly. We first investigate the overall picture of collaboration among these scholars. We want to know how the authors' attributes and structures of their collaboration networks spur one author to cooperate or not with other scholars. In this paper, we fit the ERGMs twice. In the first ERGM, we want to know how the effects brought by the authors' attributes influence the generation of ties in this coauthorship network. So we first model both the main and homophily effects of authors' attributes: the number of publications in different types, the number of citations, the most frequently used research topic and the gender in Model I. In the second

ERGM, a more comprehensive model is fitted, in which the effects of several local network structures are added (see Equation 1).

Table 1 shows the results. As indicated by the AIC, the model fit index, we find that the second model which includes the effects of both authors' attributes and networks structures has a better performance, with the AIC improving from 276,325 to 8323 (indicating the smaller the better). The effects of authors' attributes in two models, however, almost remain the same, which demonstrates that the modeling of authors' attributes is stable and reliable. Taking the network' structures into consideration thus enables a better explanation of the network's formation. Model II shows the ways in which authors' attributes and the network's structures simultaneously affect the generation of the scholarly network.

Table 1. *ERGM Results for Modeling the Coauthorship Networks among the Most Productive Authors*

	Model I		Model II	
Variables	Est.	SE	Est.	S
Main Effects				
No. of single-authored publication	0.03		0.07	
No. of first-authored publication	0.06	*	0.05	
No. of non-first-authored publication	0.06	*	0.05	***
No. of citation	0.00	*	0.00	
Most-used Topic 2(Medical IR)	0.49	*	0.45	*
Most-used Topic 3(IR Theory)	0.04		0.00	
Most-used Topic 4(IR Systems)	0.02		0.03	
Most-used Topic 5(Image-based IR)	0.06		0.06	
Gender Female	0.17		0.39	
Homophily				

Single-authored publication no. difference	-0.15		*	-0.08	*
First-authored publication no. difference	0.03			0.00	
Non-first-authored publication no. difference	-0.02			-0.01	
Citation no. difference	0.00		*	0.00	*
Same most used topic	1.85		*	1.30	***
Same gender	0.37			0.49	*
Network Structures					
Transitivity	-----------	——	——	2.46	***
Preferential attachment	-----------	——	——	0.64	***
Edges	-7.88		*	-8.90	***
Model Fit: AIC(Smaller is better)	276325			8323	

NOTES: *p<0.05, **p<0.01, ***p<0.001for a two-tailed test

We further interpret the results and examine the effects of several features mentioned above. Several points are particularly relevant to the paper's outcomes.

Main effects and homophily effects of authors' attributes. We first notice that the number of author's publications in which they did not serve as first authors has a significantly positive influence on their collaboration in both models. From the estimates we find that if keeping all the other features unchanged, the probability of gaining a new collaboration tie for one author who has one more unit of non-first-authored publication is 1.06 ($e^{0.06}$) times that of the other author in Model I and 1.05 ($e^{0.05}$) times in Model II; the probability of collaboration increases by 6%/5% when adding one unit of non-first-authored publication. Though such effect is not very strong, it verifies the pattern we observe in Figure 1A, wherein the larger-sized nodes (those authors who have produced more non-first-authored publications) tend to have a few more connections.

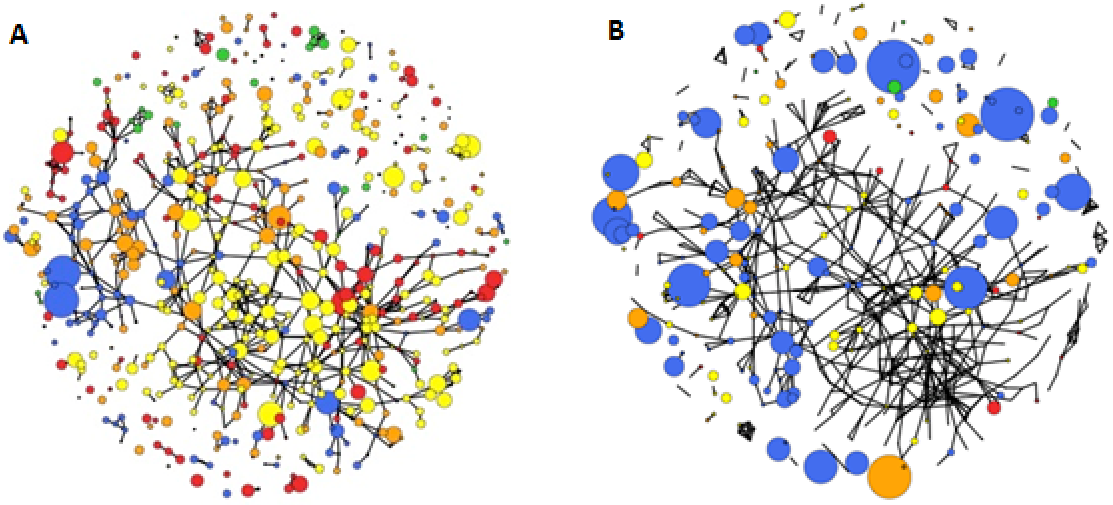


Figure 1. Plot of collaboration—(A) nodes sized by number of non-first-authored publication; (B) nodes sized by number of single-authored publication

(Note: each color represents one topic: {yellow: Topic 1 Database},{green: Topic 2 Medical IR},{orange: Topic 3 IR Theory},{blue: Topic 4 IR Systems},{red: Topic 5 Image-based IR})

It is interesting that from the perspective of homophily, the number of non-first-authored publications does not affect the authors' collaboration levels. Instead, we find the number of single-authored publications has a counter-homophily effect, in that scholars who published a similar number of single-authored articles do not tend to collaborate with each other. These significantly negative estimates indicate that there are only 86% ($e^{-0.15}$) and 92% ($e^{-0.08}$) chance in Model I and Model II, respectively, that authors will select a coauthor who has published the same or a similar number of single-authored works with him/her, compared with selecting a coauthor whose number of single-authored work is quite different. This also confirms the pattern observed in Figure 1B, in which each node stands for one scholar in the coauthorship network while the size reflects his/her number of single-authored publications, collaborative ties emerge more between nodes with different sizes, than those with similar sizes.

It is also interesting that by examining the number of citations one author has received, we find this value has no relation with the authors' collaboration at all: neither the authors with more citations tend to receive more collaboration (from the main effect perspective), nor the authors with the same or a similar number of citations are more likely to coauthor with each other (from the homophily effect perspective). While this result is beyond our expectation, it may imply that the number of citations, which is one measurement of an author's popularity (Ding and Cronin, 2011), is not a driving force for scientific collaboration. Though our measurement is targeted on authors, it may also strengthen Hsu and Huang's (2010) argument that "predicting the citation number from the coauthor number can be more reliable than predicting the coauthor number from the citation number" (p. 317).

In the "topic" related terms under "Main Effects" of both models, we compare the probability of gaining collaboration for the authors who most preferred Topic 2, 3, 4, and 5 with the baseline where those authors most used Topic 1, respectively. For example, we find that in both models, authors who liked the topic Medical Information Retrieval most significantly have a higher chance to attract new collaborators, where the probability is increased by 63% ($e^{0.49} - 1$) in Model I and 57% ($e^{0.45} - 1$) in Model II. We also observe the existence of the homophily effect in authors' research interests. The probability that authors collaborate with those who share the same research topic is 6.36($e^{1.85}$) times in Model I and 3.67($e^{1.30}$) times in Model II compared to the probability of the authors' collaborating with those not sharing their interest. When plotting the coauthorship network with nodes colored by the representing author's most used topic (shown in Figure 2), we find that there exist many same-color clusters, which demonstrates the authors' preference for same or similar topic collaboration. This finding is consistent with the conclusion of Ding (2011), that in the field of information retrieval, productive authors tend to directly collaborate with those who share the same research interests, and not directly coauthor with those who work on different topics.

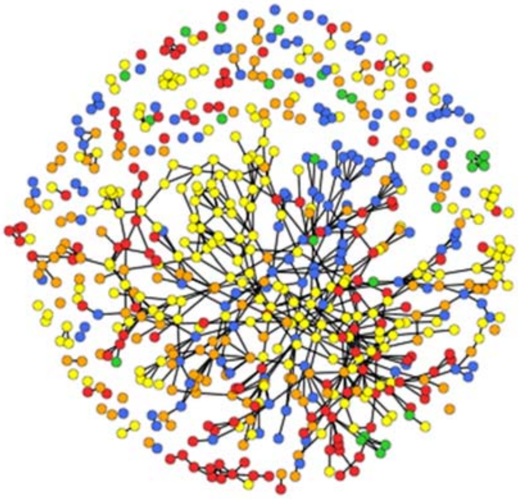


Figure 2. Plot of collaboration

(Note: each color represents one topic: {yellow: Topic 1 Database},{green: Topic 2 Medical IR},{orange: Topic 3 IR Theory},{blue: Topic 4 IR Systems},{red: Topic 5 Image-based IR})

When examining the gender distribution in collaboration, we notice that to some extent the authors' gender influences the patterns of authors' collaboration. Though neither male nor female is found to have a significant advantage in attracting collaboration, we find that there exists the homophily effect in authors' gender; scholars tend to collaborate with others who have the same gender. The probability of same-gender coauthorship is $1.45(e^{0.37})$ times in Model I and $1.63(e^{0.49})$ times in Model II compared to the probability of cross-gender collaboration. When plotting the coauthorship network with nodes colored by the representing author's gender (shown in Figure 3), we find that there exist many same-color clusters, which demonstrates the authors' preference for the same-gender collaboration.

In general, our first hypothesis has been verified that homophily does influence the formation of scientific collaboration network. Specifically, we find that the authors' productivity measured by their single publication numbers has a counter-homophily effect; the authors' impact does not bring any related homophily effect; but both the authors' research topics and their gender do bring strong homophily effect on the creation of collaboration between authors.

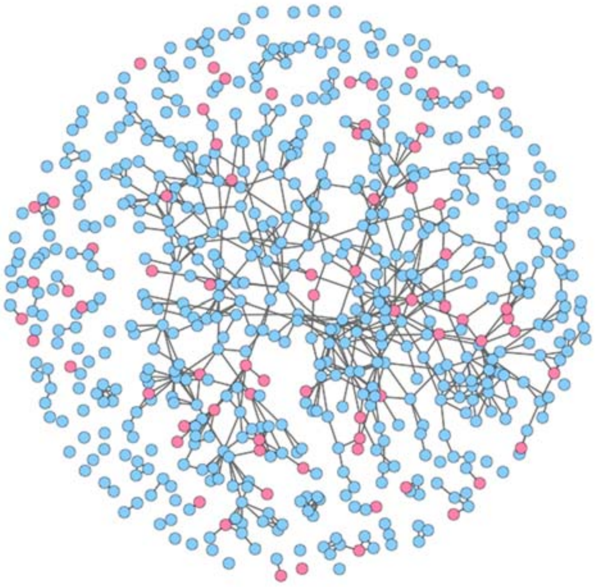


Figure 3. Plot of collaboration

(Note: each color represents one gender: {blue: male}, {pink: female})

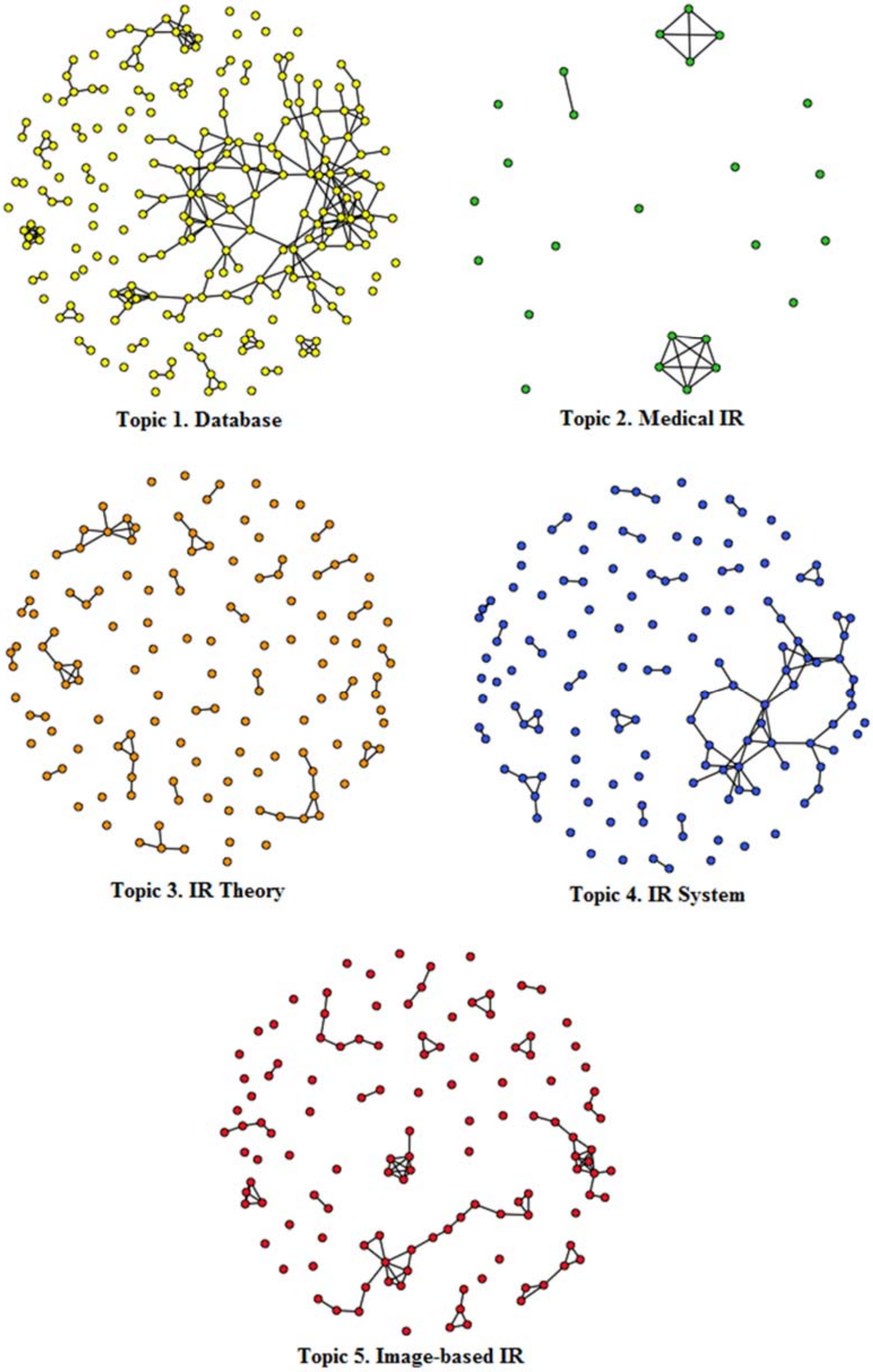
Transitivity. We see that the parameter of the transitivity in the authorship network is positive and significant, which means that the effect of network's transitivity strongly influences the authors' collaborations. The value of this parameter stands out in the fitted model, where remarkably, the probability of one author collaborating with his/her coauthors' coauthors is 11.70 ($e^{2.46}$) times the probability of not collaborating. The triangular collaboration, however, is far more likely to occur. This result conforms to Newman and Park's (2003) study point that the coauthorship network as a typical social network which has a high level of transitivity. We thus observe a few triangular clusters in the collaboration graph in Figure 2. The second hypothesis that transitivity effect plays an important role in the formation of scientific collaboration network has been verified.

Preferential attachment. From the study results, we find a significant effect of preferential attachment in this coauthorship network. The probability that one author would like to collaborate with another researcher who has more than one previous collaborator is about 1.90 ($e^{0.64}$) times than those without collaborators. A core-periphery structure thus emerges, wherein a large group of nodes (see the lower part of Figure 2) are well connected and form the “core” structure, which represents that this author preference accounts for most collaborations in the network. The other nodes are loosely connected and form the “periphery” structure, and these authors have lower levels of collaboration. The third hypothesis that the effect of preferential attachment plays an important role in the formation of scientific collaboration network has been confirmed.

ERGM Results on Each Topic-subnetwork

Figures 4 a to e show the collaboration among the authors whose core research interests are Topic 1 to 5, respectively. From the figures, it is clear that the collaboration extent and patterns within each sub-area of the information retrieval field are quite different. There are more scholars who mostly worked on Topic 1 (database), compared with those working on other topics. They have denser collaboration, and there is a big connected cluster in this subnetwork. The researchers in Topic 2 (Medical IR) are the least in number. We see two small closed clusters in this subnetwork, each of which reflects the scholars are fully connected, as any two of them have coauthored before. The scales of the other three subnetworks, formed by scholars working on Topic 3 (IR Theory), Topic 4 (IR Systems), and Topic 5 (Image-based IR) are similar. But the collaboration patterns vary. Collaboration among scholars who published most in Topic 3 is more restricted to small groups (formed by less than 10 people). Except for some pairwise and triple-wise coauthorship, however, there is a rather big and dense collaboration circle among scholars who work on Topic 4, wherein most of the authors have at least three collaborators. While the collaboration among those whose main topic is Topic 5 shows chain structures, where one author collaborates with the other one, the other author further collaborates with the third one.

Since the subnetwork of authors whose main topic is Topic 2 is too sparse, we do not fit it into the ERGM model. For the other four topic subgraphs, we use the same model to fit each collaboration network. Table 2 shows the results. We find that in the subnetwork of Topic 3 (IR Theory), the more non-first authored publications one scholar has, the more likely he/she will gain new collaborators. In the subnetwork of Topic 4, IR Systems, researchers tend to coauthor with those who have written quite different amounts of single-authored publications. From the perspective of network structure effects, we find strong effects of transitive collaboration and preferential attachment when generating coauthorships within each topic of information retrieval.



Topic 1. Database

Topic 2. Medical IR

Topic 3. IR Theory

Topic 4. IR System

Topic 5. Image-based IR

Figure 4. Plots of collaboration in each topic subnetwork

Table 2. *ERGM Results for Modeling the Collaboration on 5 Topic Subnetworks*

	Topic 1	Topic 3	Topic 4	Topic 5
Variables	Est.	Est.	Est.	Est.
Main Effects				
No. of single-authored publication	0.08	-0.01	-0.01	-0.02
No. of first-authored publication	0.00	0.07	-0.02	0.04
No. of non-first-authored publication	0.02	0.08	0.02	0.04
No. of citation	0.00	0.00	0.00	0.00
Homophily				
Single-authored publication no. difference	-0.14	-0.15	-0.10	-0.11
First-authored publication no. difference	0.06	0.09	0.07	0.1
Non-first-authored publication no. difference	-0.01	0.00	-0.02	-0.01
Citation no. difference	0.00	0.00	0.00	0.00
Network Structures				
Transitivity	2.1	2.2	1.3	2.4
Preferential attachment	1.5	1.6	0.9	0.8
Edges	-6.6	-8.2	-6.3	-7.3
Model Fit: AIC(Smaller is better)	2411	6389	8674	8219

NOTES: *p<0.05, **p<0.01, ***p<0.001for a two-tailed test

Conclusion

This paper provides a systematic analysis of scientific collaboration by comprehensively considering the simultaneous effects of scholars' characteristics, including their productivities, number of citations, their main research interests, and their gender, as well as the homophily effect on these attributes, the transitivity, and the preferential attachment of their collaboration networks by using ERGMs. The model built in this work reflects the generation of actual research collaboration, where authors' individual characteristics and the structure of their embedded network interactively determine their collaboration levels and preferences. We show concretely how each effect contributes to the collaboration simultaneously.

The major findings of this paper can be summarized as follows: the different types of publications one author has written play different roles in his/her collaborations, where the more papers he/she has collaborated with others as non-first authors, the more likely he/she will gain more collaboration; while people prefer not to coauthor with those who have produced similar numbers of single-authored publications. The number of citations does not influence authors' collaboration preference, though sharing the same research interest or having the same gender is crucial for generating collaboration between two authors. An author's tendency to form new collaborations with his/her coauthors' collaborators is strong. The more coauthors one author has, the more new collaborators he or she will attract.

From our analysis results, we find that the transitivity has the largest effect in forming this collaboration network, followed by the homophily effect of authors' research interest, the effect of preferential attachment, and the homophily effect of gender. The probability that a scholar will find a coauthor from his/her coauthors' coauthors is about 3 times of that from researchers who share the same research interest with him/her. Meanwhile, the probability for a scholar to find a coauthor from those sharing the same research topics is about 2 times of that to find one if he/she has one more previous collaborator before; it is also about 2 times of that to find a coauthor from the same gender scholars.

From a deeper examination of collaboration at each topic level of the information retrieval field, we find that Medical Information Retrieval is the topic where there is a tendency to collaborate with those authors whose main interests lie in. The number of non-first authored and single-authored publications on the topic of Information Retrieval Theory and Information Retrieval Systems, respectively, most influences the levels of within-topic collaborations. Scholars who mainly work on each topic presented herein all have a high propensity to collaborate with their coauthors' coauthors, or those who already have many collaborators. Both influences are the strongest for authors whose core topic is Information Retrieval Theory.

We could see that it is very likely for researchers who share the common coauthors to form new collaboration. Such a collaboration will result in a low search cost since the two parties of the collaboration could be introduced by the common collaborator. It can also reach high efficiency since the two parties may have more common language by sharing the same collaborator. Collaborating with coauthors' coauthors also provide a path for a certain scholar to reach others from outside domain; thus a cross-disciplinary collaboration could be formed, which has rather high innovation. We believe that more opportunities and resources should be provided to support such kind of collaboration. For example, special funding could be offered to a certain scholar to arrange some informal meetings to bring his/her coauthors together; particular forums with similar purpose may be organized in some conferences; conference organizers could even invite the coauthors of a speaker to attend the same session. By these means, more isolated scholars could talk to each other and trigger future collaboration. We also observe that well-established scholars have higher chance to attract collaborators. For a newcomer, it could be practical for him/her to connect with those popular researchers to seek for more collaboration opportunities.

We also find that it is easier for the scholars sharing the same topic or the same gender to form new collaboration. However, it is widely recognized that multidisciplinary collaborations promote innovation (Cummings & Kiesler, 2005). “Innovation opportunities going forward will be at the cusps of different disciplines” (Mitra, 2009). “Multidisciplinary projects should increase the likelihood of innovation due to their juxtaposition of ideas, tools, and people from different domains.” (Cummings & Kiesler, 2005, p. 704). The within topic collaboration may limit the innovation and creativity of the scientific products. Without taking any actions, people would be more reluctant to coauthor with those working on different topics. Some interventions should be taken to encourage cross-topic collaboration. For example, funding agencies could continue requiring the collaboration teams to be assembled by researchers from different domains. Universities could continue advocating cross-institution collaboration by increasing the credit of such projects when evaluating the participants. What is more, to make it more efficient for scholars to find the right persons to collaborate, universities could host different workshops and meetings to help scholars from different departments to get familiar with other topics; a common ground could be developed during such process. Actions should also be taken to foster the trust among collaborators from different domains. Agents should also keep exploring the way of effective formation of cross-disciplinary collaboration, for example, encouraging collaboration among scholars who come from different background but share the same coauthors. We also realize that the gender could be a barrier for initiating collaboration. For example, if restricting the collaborators to be those who have the same gender, it will be difficult for the researchers who are the minority gender in their environment. Certain policies should be implemented to encourage scholars in the majority gender to collaborate with those who have different gender, for example, when making the tenure decisions, institutions could assign more credits to the majority gender scholars who have conducted research with mix-gender teams. For the minority gender group, they are encouraged to conduct mix-gender collaboration; but the reward system will not penalize their same-gender collaboration due to the existing gender imbalance.

In this study we have selected a bunch of common authors' attributes, such as the productivity measured by different ways, the research topic, and gender, the homophily effects on these attributes, the transitivity and the preferential attachment. Though scholars in different disciplines may have different tendency to collaborate (Birnholtz, 2007) and different disciplines do have various intellectual organization (Fuchs, 1992; Whitley, 2000), we believe that in most disciplines, the collaboration networks are social networks, thus will have some attributes in common, for example, displaying a high degree of transitivity (Newman, (2001a) demonstrated in the fields of both physics and biology and medicine); the preferential attachment is also an important property in many large networks, such as collaboration networks in mathematics and neuroscience (Barabási et al, 2002); the homophily patterns have been observed in a few networks, such as collaboration networks in economics (Boschini & Sjögren, 2007). The perspectives we select to investigate collaboration are three important features in social network studies. Thus we believe the three perspectives in our study could be applied to investigate collaboration in other fields.

In this work, we apply the ERGMs in a binary coauthor network. Due to the network sparseness, unfortunately we could not apply the same procedure in the weighted network. In the future, we will apply ERGMs to the author citation networks and overlay outcomes with the coauthorship network to study whether the impact can drive the levels and preferences of scientific collaboration. In addition, we will conduct a temporal analysis to better understand the dynamic aspects of scientific collaboration. We plan to model the change of coauthorship in a time range and investigate the ways in which the authors’ individual characteristics and structure features of their embedded network dynamically influence their scientific collaboration.

References

Aghagolzadeh, M., Barjasteh, I., & Radha, H. (2012, August). Transitivity matrix of social network graphs. In *Statistical Signal Processing Workshop (SSP), 2012 IEEE* (pp. 145-148). IEEE.

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.

Boschini, A., & Sjögren, A. (2007). Is team formation gender neutral? Evidence from coauthorship patterns. *Journal of Labor Economics, 25*(2), 325-365.

Boudreau, K., Brady, T., Ganguli, I., Gaule, P., Guinan, E., Hollenberg, T., & Lakhani, K. R. (2014). A field experiment on search costs and the formation of scientific collaborations.

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

Cummings, J. N., & Kiesler, S. (2005). Collaborative research across disciplinary and organizational boundaries. *Social studies of science*, *35*(5), 703-722.

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

de Solla Price, D. J., & Beaver, D. (1966). Collaboration in an invisible college. *American Psychologist*, *21*(11), 1011.

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics*, 5(1), 187-203.

Flint, J., & Munafò, M. (2014). Schizophrenia: Genesis of a complex disease. *Nature*, *511*(7510), 412-413.

Franceschet, M., & Costantini, A. (2010). The effect of scholar collaboration on impact and quality of academic papers. *Journal of Informetrics*, *4*(4), 540-553.

Franceschet, M. (2011). Collaboration in computer science: A network science approach. *Journal of the American Society for Information Science and Technology*, *62*(10), 1992-2012.

Freeman, R. B., & Huang, W. (2014). Collaborating with people like me: Ethnic co-authorship within the US (No. w19905). *National Bureau of Economic Research*.

Goodreau, S. M., Kitts, J. A., & Morris, M. (2009). Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks*. *Demography, 46*(1), 103-125.

Hsu, J. W., & Huang, D. W. (2010). Correlation between impact and collaboration. *Scientometrics, 86*(2), 317-324.

Huang, J., Zhuang, Z., Li, J., & Giles, C. L. (2008, February). Collaboration over time: characterizing and modeling network evolution. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (pp. 107-116). ACM.

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

Katz, J. (1994). Geographical proximity and scientific collaboration. *Scientometrics*, *31*(1), 31-43.

Kraut, R., Egido, C., & Galegher, J. (1988, January). Patterns of contact and communication in scientific research collaboration. In *Proceedings of the 1988 ACM Conference on Computer-supported Cooperative Work* (pp. 1-12). ACM.

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science*, *35*(5), 673-702.

Leimu, R., & Koricheva, J. (2005). Does scientific collaboration increase the impact of ecological articles?. *BioScience*, *55*(5), 438-443.

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.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 415-444.

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.

Mitra, S. (2009, January 30). Barriers to innovation [Web log post]. Retrieved from http://www.forbes.com/2009/01/29/entrepreneur-venture-capital-technology-enterprise-tech_0130_innovate.html

Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, *69*(2), 213-238.

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

Newman, M. E. (2001b). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404-409.

Newman, M. E. (2001c). Who is the best connected scientist? A study of scientific coauthorship networks. *Phys. Rev. E*, *64*(016131).

Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5200-5205.

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

Pao, M. L. (1982). Collaboration in computational musicology. *Journal of the American Society for Information Science*, *33*(1), 38-43.

Pravdić, N., & Oluić-Vuković, V. (1986). Dual approach to multiple authorship in the study of collaboration/scientific output relationship. *Scientometrics, 10*(5-6), 259-280.

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., 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.

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.

Schilling, M. A., & Phelps, C. C. (2007). Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. *Management Science*, *53*(7), 1113-1126.

Sie, R. L., Drachsler, H., Bitter-Rijpkema, M., & Sloep, P. (2012). To whom and why should I connect? Co-author recommendation based on powerful and similar peers. *International Journal of Technology Enhanced Learning, 4*(1-2), 121-137.

Solomon, J. (2009). Programmers, professors, and parasites: Credit and co-authorship in Computer Science. *Science and Engineering Ethics*, *15*(4), 467-489.

Tang, J., Jin, R., & Zhang, J. (2008, December). A topic modeling approach and its integration into the random walk framework for academic search. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on* (pp. 1055-1060). IEEE.

Thurman, P. W., & Birkinshaw, J. (2006). Scientific collaboration results in higher citation rates of published articles. *Pharmacotherapy*, *26*(6), 759-767.

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

Wasserman, S., & Robins, G. (2005). An introduction to random graphs, dependence graphs, and p*. *Models and methods in social network analysis, 27*, 148-161.

Yu, Q., Long, C., Lv, Y., Shao, H., & He, P. (2014). Predicting Co-Author Relationship in Medical Co-Authorship Networks. *PLoS ONE*, *9*(7), e101214.

Appendix

A Review of the ERGMs Algorithm

Available upon request.