Understanding Scientific Collaboration from the Perspective of Collaborators and Their Network Structures

Chenwei Zhang[1], Yi Bu[2], Ying Ding[1]
[1]Department of Information and Library Science, Indiana University Bloomington
[2]Department of Information Management, Peking University

Abstract
Scientific collaboration is one of the key factors to trigger innovations. Coauthorship networks have been taken as representations of scholars' collaboration for a long time. This study investigates how the authors' attributes and the coauthorship network structures simultaneously influence the scientific collaboration among them. Exponential random graph models (ERGMs) are adopted in this research. We find that an author has a propensity to coauthor with the other scholar if they have different levels of productivity. We also find that the effect of network's transitivity strongly influence authors' collaboration. We demonstrate that taking the effects from both authors' attributes and the network structures into consideration helps gain a comprehensive understanding of scientific collaboration.

Keywords: Coauthorship networks; exponential random graph models (ERGMs); scientific collaboration; bibliometrics
doi: 10.9776/16470

Contact: zhang334@indiana.edu

1 Introduction

Scientific collaboration is one of the key factors triggering innovations; many scholars have investigated it from various aspects with both qualitative and quantitative methods, such as the formation and evolution of research teams, characteristics of coauthorship, and evaluation of multi-author articles. Collaboration forms a type of social network in which scholars interact with each other to collaboratively explore/solve problems in one domain or across different domains. However, previous research on scientific collaboration focuses on either scholars' individual characteristics or quantitative measurements of scholars' impact from the coauthorship network structures. Actually, it is important to consider the interplay of different authors' attributes and the social network effects to gain a holistic understanding of scientific collaboration. Coauthorship is one of the most well-documented and tangible scientific collaborations. In this poster, we analyzed a coauthorship network in the field of information retrieval (IR) using Exponential Random Graph Models (ERGMs) to see how authors' attributes, such as productivity, popularity, and their research interests, and the related homophily effects, interact with social network effects, such as transitivity and preferential attachment, to affect the authors' scientific collaboration.

Homophily is a fundamental effect in social networks. It refers to that nodes have a tendency to make connections with those that are similar with themselves in the networks. Homophily, as an important covariate effect of nodes' attributes, may cause the formation of communities in a network. Homophily has been observed in scientific collaboration. Freeman and Huang (2014) found authors with similar ethnicities had more frequent collaboration than predicted from their proportion among authors. Transitivity is also a common phenomenon in social network. It means there is a high probability of two nodes being connected if they are connected to one or more common nodes. Newman (2001) found "the probability of a pair of scientists collaborating increases with the number of other collaborators they have in common" (p. 1). Preferential attachment is another common process in social networks. It means the more existing ties one node has the more new connections it is likely to accumulate. Preferential attachment is related to the theory of cumulative advantage in science, known as Matthew Effect (Merton, 1968). Newman (2001) found the number of new publication one author gained each year increased with the number of his past collaborators, which demonstrates the existence of preferential attachment in scientific collaboration.

Our work differs from previous studies because of the comprehensive consideration of both covariate effects of authors' attributes and social network structure features to understand research collaboration, rather than examining each feature in isolation. Meanwhile, ERGMs allow us to calculate the possibilities that two authors might collaborate based on the effects of their own attributes, homophily on these attributes, transitivity, and their preferential attachment. In addition, our quantitative research is based on a relatively large network.

2 Methodology

2.1 Data

Papers and their corresponding citations were harvested from Web of Science from the years 1956 to 2014 in IR. IR is a transdisciplinary area, with most collaboration formed by two or three authors (Franceschet, 2011). Unlike disciplines having a large list of coauthors, such as biomedicine and high-energy physics (Cronin, 2001), every co-author in one publication in IR has a significant level of involvement in the collaboration. The coauthorship in IR can be an effective source for studying scientific collaboration. We referred to Ding (2011) for a list of query terms. The dataset contains 59,162 authors publishing 20,359 papers, in which there are 558,498 references.

2.2 Coauthorship Networks and Authors' Attributes

We first ranked all the authors by each author's publication number. Initially we wanted to select top 500 productive authors. Since authors from the 447th to 631th all have six publications, we included all of them. We collected the number of authors' publications (productivity) and citation received (popularity). We used Author-Conference-Topic (ACT) model by Tang, Jin and Zhang (2008) to extract the authors' research topic distribution. Sixty-six authors have equal weights for multiple topics and their top two research interests could not be decided. After removing them, we generated a coauthorship network among the most productive 565 authors. Each author represents one node in the network. We focused only on the collaboration with different coauthors, without considering the frequency of collaborations. So if two authors have collaborated before, there is a tie between them; otherwise, there is no tie.

2.3 Exponential Random Graph Networks

We applied ERGMs (Robins, Pattison, Kalish, & Lusher, 2007) to model the scientific collaboration. ERGMs are the state-of-the-art approaches for modeling how network is formed, i.e., how each tie is created. The basic assumption of ERGMs is that there is a family of networks which have the same number of nodes as the current coauthorship network (only one observation); the possible ties among authors in the network are treated as random variables; every collaboration tie between any two authors can be explained by any network structures features and the nodes' attributes. The probability of observing the current network (w) is

$$\Pr(W = w|X) = \frac{\exp\{\theta^T g(w, X)\}}{K}$$

where W is a random network, X the authors' attributes, $g(w, X)$ the features including covariate effects of authors' attributes, such as the number of papers he/she published, the number of citations he/she received, the top two topics he/she was interested in, and the corresponding homophily; and network structure effects, such as transitivity and preferential attachment, θ a vector of parameters, which estimates the effects of these features on network formation, and K the normalizing factor that ensures the probabilities sum to 1. Here we fitted ERGMs twice. Model I only focused the effects of authors' attributes separately, while Model II added effects of several local network structures.

3 Preliminary Results

Figure 1 shows the collaboration among the most productive authors, in which the size of the vertex represents the number of publications this author published; the color is based on the author's research interest, which has the highest probability among his/her research topic distribution. Table 1 shows the ERGMs results.

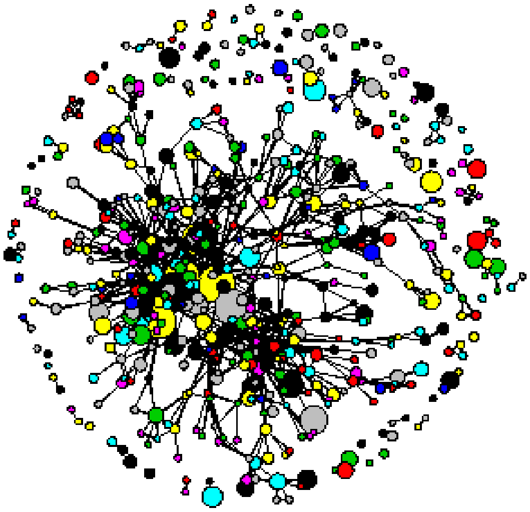

Figure 1. Coauthorship network among the most productive authors

	Model I			Model II		
Variables	Est.	SE		Est.	SE	
Main Effects						
Number of paper published	0.07	0.01	***	0.05	0.01	***
Number of citation received	0.00	0.00		0.00	0.00	
Homophily						
Publication number difference	-0.11	0.01	***	-0.09	0.02	***
Citation number difference	0.00	0.00		0.00	0.00	
Same most used topic	-0.13	0.17		0.05	0.28	
Same second most used topic	0.43	0.14	**	0.44	0.23	
Network Structures						
Transitivity				2.58	0.09	***
Preferential Attachment				0.20	0.19	
Edges				-7.30	0.20	***

Table 1. ERGM results for modeling the coauthorship networks among the most productive authors

3.1 Homophily Effect

From the significantly negative coefficient of “publication number difference,” we find researchers do not tend to coauthor with those having similar number of publications with them; instead, they prefer those having different level of productivity. This is observed in Figure 1 where there are many connections between larger-sized and smaller-sized nodes. By examining one author’s number of, it is surprising that this value has no relation with the authors’ collaboration. This may imply the number of citations, which is one measurement of an author’s popularity (Ding & Cronin, 2011), is not a driving force for scientific collaboration. Figure 1 shows nodes with different colors are well mixed, which means scholars with different research interests collaborated with each other. By comparing the coefficients of authors’ homophily patterns in research topics, we do not see a consistent and significant homophily effect, which means research interest similarity does not necessarily affect which scholars one author will collaborate with. This reflects when selecting collaborators, one scholar does not take the research topic difference as his/her top priority.

3.2 Transitivity Effect

We notice the coefficient of transitivity in the network is positive and significant, which means the effect of network's transitivity strongly influences authors' collaboration. The probability of one author collaborating with his/her coauthors' coauthors is much higher than that of not collaborating. The triangular collaboration is very likely to occur. These transitive structures are also reflected in Figure 1. This result conforms to Newman and Park (2003)'s research that coauthorship network is a typical social network, which has a high level of transitivity.

3.3 Preferential Attachment Effect

The results show there does not exist a significant effect of preferential attachment in this coauthor network. Therefore an author who already has many collaborators will not necessarily attract more coauthorship than the other one with fewer previous collaborators.

4 Conclusions

This research reports how authors' attributes and social effects from their coauthorship network structures affect their collaboration simultaneously. The findings are informative: there is a propensity for authors to collaborate with each other if they have different levels of productivity; there is also a strong tendency for an author to form new cooperation with his/her coauthors' collaborators. In the future, we will extend the study to several other disciplines. We will focus on the topic diversity of coauthors and their collaboration tendency. We will also apply ERGMs to the author citation networks and overlay with the coauthorship network to study whether the impact will drive the scientific collaboration.

5 References

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices?. *Journal of the American Society for Information Science and Technology,* 52(7), 558-569.

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

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

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*.

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

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.

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

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.