International Journal of Modern Physics C
Vol. 25, No. 5 (2014) 1440014 (8 pages)

DOI: 10.1142/S0129183114400142

World Scientific
www.worldscientific.com

Homophily versus preferential attachment: Evolutionary mechanisms of scientific collaboration networks*

Zhen-Zhen Wang† and Jonathan J. H. Zhu‡

Web Mining Lab, Department of Media and Communication
City University of Hong Kong
Tat Chee Avenue Kowloon, Hong Kong SAR
†*wzzjasmine@gmail.com*
‡*j.zhu@cityu.edu.hk*

Received 30 June 2013
Accepted 30 September 2013
Published 4 December 2013

Homophily and preferential attachment are among the most recognized mechanisms of network evolution. Instead of examining the two mechanisms separately, this study considers them jointly in a scholarly collaboration network. Specifically, when a new scholar enters a field, how does he/she choose the first collaborator from the pool of available scholars? We find that new scholars tend to collaborate with someone who works in the same institution (which is called constrained acceptance), shares similar specialty interests (active choice), or has already worked with many collaborators (random action). We view constrained acceptance and active choice as supporting evidence for homophily (because similarity is attractive) and random action as supporting evidence for preferential attachment (because popularity is attractive). As such, both homophily and preferential attachment affect the evolution of collaboration networks. Furthermore, the influences vary over time with random action, constrained acceptance, and active choice taking turns to act the dominant force at the beginning, middle and later phases of the evolution process, respectively.

Keywords: Constrained acceptance; active choice; random action; institutional homophily; specialty homophily.

PACS Nos: 11.25.Hf, 123.1K.

1. Introduction

Network studies have focused primarily on mechanisms in the evolution of networks. Preferential attachment[1] is among the most famous mechanisms. It argues that new nodes will attach preferentially to already well-connected nodes, i.e. popular nodes which have high degree centrality. If preferential attachment implies that popularity is attractive, similarity might be another dimension of attractiveness[2] that similar

*The study was supported in part by Hong Kong Research Grants Council (GRF CityU 154412) and City University of Hong Kong (SRG 7002652).

nodes have a higher probability to get connected to each other. The phenomenon is known as homophily.[3]

Instead of solely emphasizing either of the two mechanisms, this study examines them jointly in a scholar collaboration network. Specifically, we examine whether new scholars tend to collaborate with someone who works in the same institution (which is called constrained acceptance), shares similar specialty interests (active choice), or has already worked with many collaborators (random action). We view constrained acceptance and active choice as supporting evidence for homophily, and random action as supporting evidence for preferential attachment. Our study indicates that the three factors contribute differently to link formation at different stages of the network evolution.

In terms of methodology, this study responds to Papadopoulos *et al.*'s critics[2] of indirect validation: When examining network evolution mechanisms, most studies simply compare parameters between simulated and real networks, that only validates the consequence of mechanisms, not mechanisms *per se*. To conduct a direct validation, we adopt a randomized null-model approach which is popular in studies of real networks (e.g. see Opsahl, Colizza, Panzarasa and Ramascos study[4]), and examine the mechanisms locally (at link level), rather than globally (at network level). By doing so, we cast network evolution as a generative phenomenon. It allows us to get closer to the real meaning of mechanism.

2. Hypotheses

2.1. *Preferential attachment and random action*

The term preferential attachment was originally used by Barabsi and Albert,[1] but the richer get richer principle it describes can be traced back to Yule process,[5] cumulative advantage[6] and the Mathew effect.[7] This mechanism is deemed to be responsible for the commonly observed power-law distribution of node degree in real networks, e.g. Internet[8] and protein network.[9] Preferential attachment has also been examined and confirmed in scholar collaboration network (e.g. Abbasi, Hossain and study[10]; Jeong, Nda and Barabsi's study[11]; Newman's study[12]).

Barabsi who makes preferential attachment a famous term is among the first to try to understand the origin of preferential attachment. Without going into any particular context, he compares preferential attachment with homophily, and thinks the former is a consequence of random actions, while the latter requires human agency. He explains the randomness as follow:

> . . .randomly select a link in a directed network, for example the links of the World Wide Web that point to a document; then connect the new node to the selected links target. The more connected nodes have an advantage here, as the chance that a new node connects to them is proportional to their degree.[13]

Barabsi's interpretation stays at a high level of abstraction. If we translate his interpretation in our particular context, we need to imagine a young student who is

ignorant and confused when beginning his academic career. He might talk to a scholar that he randomly met: Can you introduce some mentor to me? Then the scholar might immediately think of his/her collaborator. In such a situation, an existing scholar who has more collaborators has a higher chance of being referred to. Of cause, in real world, the young student will have a higher chance of successfully building connection with the recommended scholar, if his situation meets certain conditions, like working in some specialty, or in the same institution with the scholar. The randomness may not work exactly like the story above, but we cannot deny that when a new comer enters a field, his/her first choice of partners more or less bears some randomness.

Therefore, we propose Hypothesis 1 that is about random action.

H1: A new scholar has a higher probability to build a collaboration relationship with an existing scholar who has a lot of collaborators, than with an existing scholar who has relatively less collaborators.

2.2. *Homophily*

Homophily is the principle that a contact between similar people occurs at a higher rate than among dissimilar people.[3] The principle has been validated in various relationships, e.g. marriage,[14] friendship,[15] discussing partners,[16] task group[17] and founding teams.[18] The pattern seems to be robust across various relation types.

Different from preferential attachment, homophily is a multi-dimensional concept. Lazarsfeld and Merton[19] differentiated two types of homophily: (1) status homophily, including ascribed characteristics like sex, age or ethnicity, and acquired characteristics like education or occupation, and (2) value homophily, including various internal states, like attitude, belief or aspiration. Some characteristics, like race and gender, are very powerful in structuring various networks, while others, like occupation and attitudes are specific only to certain types of networks.[3]

Due to availability of data, we cannot consider all characteristics that might contribute to homophily. Instead, we try to capture two main factors that forge birds of a feather in the academic world.

2.2.1. *Constrained acceptance and active choice*

The first factor is institution homophily. Scholars from the same institution naturally have higher chance to know each other. There is a tendency of scholars to select collaborators within rather than across institutional boundaries.[20] Institution homophily works as a constraint. That is why scholars associate the term institution with terms like boundary or restriction. If scholars always live within one institution, and never make collaboration outside, it is more appropriate to view it as a constrained acceptance, which reflects the power of hierarchy or structure, rather than the will of the agency.

Thus, we propose Hypothesis 2 concerning constrained acceptance:

H2: A new scholar has a higher probability to build a collaboration relationship with an existing scholar who is affiliated to the same institution with him/her, than with someone who is affiliated to a different institution.

The second factor is specialty homophily. Although scientists collaborate to complement the knowledge and skill they lack of, they tend to choose collaborators within their own specialty area, because the shared norms of research practice or mutual interest bring them together.[20] Specialty homophily represents the active choice of human agency. The birds flock together because of their common feather-shared interest or research practice, not because of some external power. Some may argue that the specialty could also work as a constraint, like people sharing the same academic interest may go to the same conference and have a higher chance of knowing each other and becoming friends. When the internal attributes, like interest or attitude, are externalized as some habit or routine, they do structure people's behavior, which make them look like constraint. However, the origin of the habit or routine can be controled and changed by the human agency. With this consideration in mind, we think specialty homophily reflects the active choice of scholars.

Therefore, we propose Hypothesis 3 concerning active choice:

H3: A new scholar has a higher probability to build a collaboration relationship with an existing scholar who shares more similarity with his/her specialty, than with an existing scholar who has a dissimilar specialty.

The linking behavior of new scholars might be forged by the combination of the three factors. But, we do not know which factor dominates. Moreover, will the dominance change at different stages of network evolution? We propose our research question:

RQ1: Which process (random action, constrained acceptance, or active choice) dominates at different stages of the collaboration network evolution?

3. Method

3.1. *Data*

In this study, we focus on the dynamics of scholar collaboration network in the field of communication.

ISI Web of Science has been used to retrieve the data in 2013. From 2000 to 2011, the database provided journal citation report which included lists of journals in different disciplines. The 12 lists for communication field were summarized into a list of 83 journals. According to the list, we searched articles published from 1970 to 2012 in communication field. Document type was limited to research articles. Document-level information from 37 868 relevant articles was retrieved, including author(s), author address, article title, abstract and author keywords.

3.1.1. *Author disambiguation*

To solve the problem of homonymy, we use an disambiguation algorithm based on multi-similarity indicators.[21,a] Through disambiguation, authors are linked to their respective corpus of work. 71 579 authors are identified as 41 119 scholars who publish at least one paper in the field communication from 1970 to 2012. A collaboration network is extracted from the coauthorship information.

3.1.2. *Attachment behavior identification*

During evolution of a collaboration network, links can happen: (1) between new scholars (that is, scholars who publish their first paper in the next period) and already existing scholars (2) among new scholars and (3) among existing scholars. The so-called attachment refers to the first type of links. There are 14 135 (34%) scholars who enter the field of communication by collaborating with at least one existing scholar. We focus on their attachment behavior.

3.1.3. *Randomized null-link matching*

Most studies of preferential attachment adopt an indirect validation approach that we have criticized. As revealed by Papadopoulos *et al.*,[2] solely examining nodes degree distribution cannot confirm the function of preferential attachment, because a mechanism mixing homophily and preferential attachment may lead to exactly the same degree distribution. Abbasi *et al.*[10] use a direct validation: they calculate the correlation between existing authors degree centrality and the numbers of new authors attracted to them in the following year. However, because of the one-to-many relationship embed in the model, their approach can only be used to examine preferential attachment, and naturally eliminates homophily which is based on pairwise-comparison. To direct validate and compare the proposed two mechanisms, for each of the 14 135 scholars, we randomly pick a null-link to match with their real-link. Specifically, we randomly pick up an existing author who has no collaboration relationship with the targeting author, and who has entered the network before the targeting author. Then we build a null-link between them, and compare the characteristics of the null-link and real-link.

3.2. *Measures*

3.2.1. *Random action*

To measure the effect of random action, we calculate existing authors degree, i.e. number of collaborators, at the time when the collaboration link is formed.

[a]Technical note of author disambiguation will be provided upon request.

3.2.2. *Constrained acceptance*

Based on the author address provided by Web of Science, we could determine whether two linked (real-link and null-link) authors are affiliated to the same institution. We use 1 to code for yes and 0 for no.

3.2.3. *Active choice*

Because we focus our investigation in publications from the field of communication, most involved authors are also affiliated to communication-related department. We cannot find any direct and discriminant specialty attributes from the information provided by Web of Science, so we trace scholars specialty from their publications. We combine the title, abstract and keywords of each paper, and calculate the cosine similarity between papers to determine the similarity of specialty between scholars. In each linked pair: for the new author, we use his/her first paper, the one in which the author collaborates with the existing author; for the existing author, we use his/her published paper which is the most similar to the targeting paper.

4. Results

Using logistic regression, we examine our hypotheses. Table 1 shows that institution homophily, specialty homophily and existing authors degree all significantly contribute to the formation of coauthor link. Thus H1, H2 and H3 are supported. Random action, constrained acceptance and active choice jointly influence the evolution of scholar collaboration network. We use dominance analysis[22] to determine the specific proportion of contribution of factors. From the last column of Table 1, we can see that, overall, active choice (measured by specialty homophily) seem to be the dominant process which contribute 38% to the explained R square, following by random action (measured by degree, 36%) and constrained acceptance (measured by institution homophily, 27%).

From Table 2 and Fig. 1, we can see that, at the beginning, the evolution of the collaboration network is dominated by random action. But its influence keeps falling off with time passing by. Active choice is the weakest at the beginning, but it keeps rising and finally becomes the dominant process. The influence of constrained acceptance exhibits some fluctuation, but the overall trend is falling off.

Table 1. Logistic regression for collaboration link.

	B	S.E.	Sig.	Contribution
Constrained acceptance	8.811	1.001	0.000	0.210 (27%)
Active choice	21.935	0.582	0.000	0.295 (38%)
Random action	4.222	0.098	0.000	0.281 (36%)
Constant	−4.925	0.097	0.000	
Nagelkerke's R^2	0.786			100%

Table 2. Dominance analysis for collaboration link across time.

Time period	Constrained acceptance	Active choice	Random action	Nagelkerke's R^2	New comers
1971–1980	0.321 (39%)	0.101 (12%)	0.392 (48%)	0.814	224
1981–1990	0.377 (47%)	0.108 (13%)	0.322 (40%)	0.807	1136
1991–2000	0.263 (33%)	0.241 (30%)	0.302 (37%)	0.806	2725
2001–2012	0.163 (21%)	0.347 (44%)	0.278 (35%)	0.788	10050

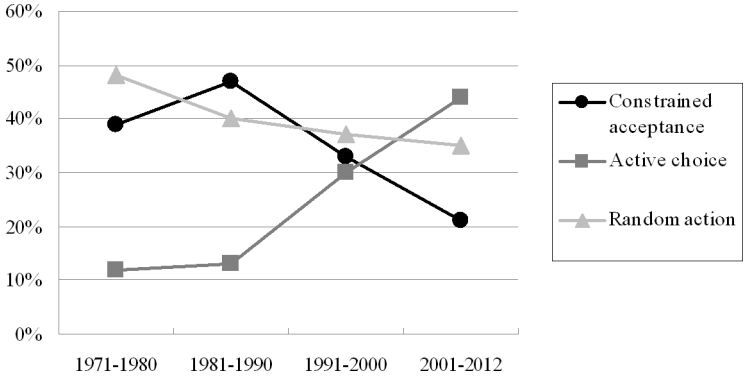

Fig. 1. Dominance analysis for collaboration link across time.

5. Discussion

In this study, we find that homophily and preferential attachment both influence the first collaboration choice that new scholars make. Specifically, a new scholar tend to collaborate with an existing scholar who works in the same institution (constrained acceptance), shares similar specialty interest (active choice) and already has many collaborators (random action). Although random action is the dominant factor at the beginning of the network evolution, active choice keeps rising and finally dominates the formation of the collaboration link. Our findings have profound implications, specifically to the scholar collaboration network and generally to the network analysis.

The active choice requires initiatives of human agency, which implies rationality. Scholars are believed to be more rational than average people. It is good to see the evidence of rationality in our scholar collaboration network. However, the phenomena may not hold in other networks. For nonsocial networks, like Internet or protein network, nodes do not possess rational thinking, thus they may exhibit totally different linking pattern. We will further investigate and make comparison between networks.

Acknowledgment

The study was supported in part by a GRF grant (CityU 154412) from the Hong Kong SAR Research Grants Council.

References

1. A.-L. Barabási and R. Albert, *Science* **286**, 509 (1999).
2. F. Papadopoulos, M. Kitsak, A. M. Serrano, M. Boguna and D. Krioukov, *Nature* **489**, 537 (2012).
3. M. McPherson, L. Smith-Lovin and J. M. Cook, *Annu. Rev. Sociol.* **27**, 415 (2001).
4. T. Opsahl, V. Colizza, P. Panzarasa, and J. J. Ramasco, *Phys. Rev. Lett.* **101**, 168702 (2008).
5. G. U. Yule, *Philos. Trans. R. Soc. London, Series B*, Containing Papers of a Biological Character **213**, 21 (1925).
6. D. J. Price, *Science* **149**, 510 (1965).
7. R. K. Merton, *Science* **159**, 56 (1968).
8. A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi and G. Caldarelli, *Phys. Rev. E* **74**, 036116 (2006).
9. E. Eisenberg and E. Y. Levanon, *Phys. Rev. Lett.* **91**, 138701 (2003).
10. A. Abbasi, L. Hossain and L. Leydesdor, *J. Inform.* **6**, 403 (2012).
11. H. Jeong, Z. Neda and A.-L. Barabsi, *Europhys. Lett.* **61**, 567 (2003).
12. M. E. J. Newman, *Phys. Rev. E* **64**, 205102 (2001).
13. A.-L. Barabási, *Nature* **489**, 507 (2012).
14. M. Kalmijn, *Annu. Rev. Sociol.* **24**, 395 (1998).
15. D. B. Kandel, *Am. J. Sociol.* **84**, 427 (1978).
16. P. V. Marsden, *Am. Sociol. Rev.* **52**, 122 (1987).
17. J. M. McPherson and L. Smith-Lovin, *Am. Sociol. Rev.* **52**, 370 (1987).
18. M. Ruef, H. E. Aldrich and N. M. Carter, *Am. Sociol. Rev.* **68**, 195 (2003).
19. P. L. Lazarsfeld and R. K. Merton, *Friendship as a Social Process: A Substantive and Methodological Analysis Freedom and Control in Modern Society* (Van Nostrand, New York, 1954), pp. 18–66.
20. T. S. Evans, R. Lambiotte and P. Panzarasa, *Am. Sociol. Rev.*, arXiv:1006.1788.
21. T. Gurney, E. Horlings and P. van den Besselaar, *Scientometrics* **91**, 435 (2012).
22. R. Azen and N. Traxel, *J. Educ. Behav. Stat.* **34**, 319 (2009).