Dynamics of Preferential Attachment in All Scientific Disciplines Citation Networks

Juyoung, An
juyan@iu.edu

Yi, Bu
@iu.edu

Yong-yeol, Ahn
@iu.edu

2017-Jul-24

Contents

This is some preamble text that you enter yourself.

1 Introduction

Citation analysis is a process that deals with data which describe formal patterns of scholarly interaction and consumption (Cronin & Overfelt, 1994) (p.20). In other words, to analyze a citation network of a scholarly domain could, to some extent, examine the domains structure and system including scholars and their publications. As a network of complex system (in this case, topology of scholarly communication in a domain), a citation network is identified as a scale-free network; such phenomenon is called preferential attachment (Barabási, Albert, & Jeong, 1999). Preferential attachment is one of mechanism which rule the growth of a large networks such as genetic networks and the World Wide Web (WWW), and denotes the tendency that when there is a new vertex added in a network the vertex attaches preferentially to already well-connected vertex. More briefly, the mechanism forms a basis for a rich-gets-richer phenomenon which widely occurs in real networks (Barabási & Albert, 1999).

In the field of science of science, preferential attachment has been explored in coauthorship and citation networks (Milojević, 2010)(M. Wang, Yu, & Yu, 2009). To explain a scholarly network of a domain tells us how equally the domains scholarly acts are conducted. Especially, given the fact that a domains citation network shows academic productivity of the domain (Eom & Fortunato, 2011), the difference between domains tendencies of preferential attachment and the dynamics of preferential attachment are important features that should be considered when academic productivity of a domain is detected. A large number of studies have successfully detected the preferential attachment in these real scholarly networks. However, these studies have at least three drawbacks. On the one hand, most of these researches have simply detected preferential attachment on a limited number of scholarly domains. For instance, two domains were employed in (Barabâsi et al., 2002)s study while only one single domain has been used in (Milojević, 2010). The limited number of domains leads to some uncertain to the phenomena of preferential attachment in various disciplines. On another hand, seldom studies have focused on the preferential attachment as a static process instead of a dynamic one, which prevents us to understand the development and evolution of preferential attachment in a long run. Essentially, the preferential attachment might have some delay effects in certain domains, as pointed out by (Sinha et al., 2015); to understand such characteristics of preferential attachment could be crucially important because it also facilitates the comparison among domains. On the other hand, even equality of citation behavior is critical to uncover academic characteristic of a domain, the most previous works has concentrated on theoretical modeling, not empirical exploration. Therefore, in this paper, we are going to detect the preferential attachment in various domains that nearly cover all scientific disciplines from a dynamic perspective on citation networks.

This paper is outlined as follows. The related studies of this research are shown in Section 2. The dataset we use and the method we propose are detailed in Section 3. The results of the dynamic change of preferential attachment among various disciplines are provided in Section 4. Some deeper discussions based on our results are illustrated in Section 5. Finally, the conclusion and limitations of this study are listed in Section 6.

2 Related Studies

2.1 Adopting Netowork Theories on Citation Networks

In citation networks, the distribution of degree has been received most attention since (Redner, 1998) conducted an empirical study to identify the distribution of the number of citations. He built the two citation networks based on a large dataset and confirmed that the degree distribution of the citation networks followed a power-law distribution. Additionally, the tendency of decrease of citation according to aging of a paper was also detected. Meanwhile, (Laherrere & Sornette, 1998) suggested stretched exponential distribution that is similar to power-law distribution, but more comprehensive then power-law distribution. They analyzed the citation of the 1,120 most cited physicists over 16-year time span and identified the fat-tail distribution of the number of citations they had got. (Redner, 2004) revisited the degree distribution of citation network by analyzing publication data from Physical Review journals for 110-year period. He concluded that preferential attachment is a core mechanism forming the power law distribution of the citation network while there is a negative correlation between the probability of being cited for a paper and the age of the paper. Whereas the existing studies have simply dealt with in-degree of citation networks, (Vazquez, 2001) identified the distribution of out-degree of a citation network. In other words, the degree of a node means the number of references a paper contains, so the distribution is various according to a journal the paper had published in. (Franceschet, 2012) also explored the distribution of out-degree of a journal citation network, and it is revealed that journal citation network has a long tail distribution for both in-degree and out-degree. At the moment, the other concepts derived from network science such as community structure, robustness, and small world have been adopted to understand a journal citation network. (Peterson, Pressé, & Dill, 2010), for example, divided citation behavior into direct and indirect according to the route of citation to predict the behaviors with the distribution of the citation made by the two ways. The results displayed that there were different patterns of distribution according to the number of citations, and the highly-cited scientists featured a low power-law exponent in the distribution.

As allude to above previous works, the other well-known characteristic in a citation network is the age of a paper, in which (Klemm & Eguiluz, 2002) put stress on the negative correlation between the rate of citations a paper receives and increasing age of the paper to make a model for growing network. They assumed that researchers tend to not consider outdated contents of old papers or to cite a review paper talking about an old paper instead of citing the old paper. Their proposed model includes the three important features of scale-free networks: power-law distribution for the degree, preferential attachment, and negative correlation between age and attachment rate.

Comparison study on citation networks of different disciplines has also been concentrated on identifying degree distribution of the networks. (Radicchi, Fortunato, & Castellano, 2008) dealt with deviation of the number of citations among various scientific disciplines, including agricultural economics and policy, allergy, anesthesiology, astronomy and astrophysics, biology, computer science, developmental biology, engineering, hematology, mathematics, microbiology, neuroimaging, physics, and tropical medicine, to rescale the number of citation as an objective measure of scientific impact by controlling the deviation. He found that the deviation exists not only in different disciplines but also distinct time of a same discipline. (Vieira & Gomes, 2010) analyzed citation networks from five fields (biology, biochemistry, chemistry, mathematics, and physics) to identify the distribution of citations with the relationship between article features– the number of co-authors, addresses, pages, references and impact factor– and citation enhancement per domain. (Albarrán, Crespo, Ortuño, & Ruiz-Castillo, 2011) compared degree distributions of citation networks in 219

sub-fields categorized by Web of Science database. The patterns of degree distribution are so different from sub-fields that power law existed in 62% of the total number of articles in their study. (Thelwall & Wilson, 2014) extracted the articles including ten years of citation data from the 20 Scopus subjects to verify whether the distribution of citations fits with power law or not. The result of their research showed that the degree distribution of some domains citation networks followed power law distribution while that of the others followed the lognormal and the hooked power law distribution. The difference of degree distribution among the scientific fields is supported later by (Brzezinski, 2015) who also used dataset drawn from Scopus and identified that about the half of the fields do not have power law distribution in their corresponding citation networks.

2.2   Preferential Attachment in Scholarly Networks

In the field of science of science, six kinds of scholarly networks have been widely explored, citation network, co-citation network, bibliographic coupling network, coauthorship network, coword network, and topical network (Yan & Ding, 2012). The phenomenon of preferential attachment has mainly been detected in coauthorship and citation networks. In co-authorship network, preferential attachment refers to that the more existing ties one node has, the more the number of new connections it is likely to accumulate. Preferential attachment is related to the theory of cumulative advantage in science, known as the Matthew effect (Merton et al., 1968)(Price, 1976). It has been used to explain observed empirical phenomena, such as the distribution of the sizes of cities (Simon, 1955), the number of citations received by learned or high status publications (Price, 1976), and the number of links to pages on the World Wide Web (Barabási et al., 1999). The preferential attachment process often generates a long-tailed distribution following a Pareto distribution or power law in its tail (Zhang et al., 2016), a phenomenon that has been extensively demonstrated in collaboration networks. For example, (Newman, 2001) found that the number of new collaborations one author gains each year increases with the number of the authors past collaborators. (Barabâsi et al., 2002) illustrated that in two collaboration networks in mathematics and neuroscience, both new and existing scholars collaboration with others showed preferential attachment. (Milojević, 2010) analyzed the collaboration modes of authors with different statues in nanoscience. Distinguishing authors by the number of collaborators they had, she studied both collaborations by new authors with former authors, and among existing authors without previous collaborations. She found that authors with more than twenty collaborators benefit from preferential attachment when forming new co-authorships. (Sinatra, Deville, Szell, Wang, & Barabási, 2015) found that core physics showed less tribal nature during its development and hinted that the preferential attachment also decayed. Zhang, Bu, Ding, and Xu (forthcoming) have found an obvious preferential attachment phenomenon in the field of information retrieval. Using Exponential Random Graph Models, they pointed out that the probability that one author would like to collaborate with another researcher who has more than one previous collaborator is over twice than those without collaborators. A core-periphery structure also emerges by their visualizations.

As for citation networks, several previous related studies have successfully detected the phenomena of preferential attachment. In (Jeong, Néda, & Barabási, 2003), a partial citation network of Physical Review Letters is analyzed as an example of real evolving network. They declared that the citation network grows base on preferential attachment. (M. Wang, Yu, & Yu, 2008) explored the preferential attachment mechanism from a theoretical level and focused on the dynamic change of preferential attachment in citation networks. However, they failed to concentrate more on an empirical level. (Redner, 1998) has explored the distribution of citation network and

found the preferential attachment phenomena there. By focusing on citation network in the field of computational linguistics, (Radev, Muthukrishnan, & Qazvinian, 2009) used some basic measurements in network theory such as diameter, average shortest path, power law, and clustering coefficient. Wang, (M. Wang et al., 2009), have similarly identified the preferential attachment in actual citation networks to characterize the influences of papers age on the attachment property. The preferential attachment, the important mechanism of citation networks has been mostly used to predict citation frequency in terms of building a prediction model. (D. Wang, Song, & Barabási, 2013) included preferential attachment as one of three fundamental mechanisms which play critical roles in building citation history in a scholarly domain. The other mechanisms are age and novelty (importance) of a paper and the 3 mechanisms are used as features to predict future citations. Like this, preferential attachment might change during different periods of time. For instance, in some domains, like core physics, the preferential attachment was found to decay (Sinatra et al., 2015) as time goes by. Basically, understanding the dynamic change from a temporal perspective is crucial because it helps us to better understand certain domains by figuring out its knowledge domain maps (Bu, Liu, & Huang, 2016). Moreover, seldom of the previous researchers only examined the preferential attachment in one or two academic domains, in which we cannot compare the preferential attachment between multiple disciplines, especially between natural science and social science.

To compare the trend of equality in citation networks among the representative disciplines which exist in the world, the present study scrutinizes preferential attachment of citation networks built by using data extracted from articles published in journals which represent the disciplines. In addition, with advanced tool to visualize the result of analysis, the various characteristics of citation detected based on properties of a network such as degree distribution will enrich the depiction of citation topology in each discipline.

3 The second section

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Refer again to **??**. (?, ?) Note also the discussion on page **??**

3.1 Title of the first subsection of the second section

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

References

Albarrán, P., Crespo, J. A., Ortuño, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics*, *88*(2), 385–397.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, *286*(5439), 509–512.

Barabási, A.-L., Albert, R., & Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, *272*(1), 173–187.

Barabâsi, 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.

Brzezinski, M. (2015). Power laws in citation distributions: evidence from scopus. *Scientometrics*, *103*(1), 213–228.

Bu, Y., Liu, T.-y., & Huang, W.-b. (2016). Maca: a modified author co-citation analysis method combined with general descriptive metadata of citations. *Scientometrics*, *108*(1), 143–166.

Cronin, B., & Overfelt, K. (1994). The scholar's courtesy: A survey of acknowledgement behaviour. *Journal of Documentation*, *50*(3), 165–196.

Eom, Y.-H., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PloS one*, *6*(9), e24926.

Franceschet, M. (2012). The large-scale structure of journal citation networks. *Journal of the Association for Information Science and Technology*, *63*(4), 837–842.

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

Klemm, K., & Eguiluz, V. M. (2002). Highly clustered scale-free networks. *Physical Review E*, *65*(3), 036123.

Laherrere, J., & Sornette, D. (1998). Stretched exponential distributions in nature and economy:fat tails with characteristic scales. *The European Physical Journal B-Condensed Matter and Complex Systems*, *2*(4), 525–539.

Merton, R. K., et al. (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 Association 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.

Peterson, G. J., Pressé, S., & Dill, K. A. (2010). Nonuniversal power law scaling in the probability distribution of scientific citations. *Proceedings of the National Academy of Sciences*, *107*(37), 16023–16027.

Price, D. d. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the Association for Information Science and Technology*, *27*(5), 292–306.

Radev, D. R., Muthukrishnan, P., & Qazvinian, V. (2009). The acl anthology network corpus. In *Proceedings of the 2009 workshop on text and citation analysis for scholarly digital libraries* (pp. 54–61).

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*(45), 17268–17272.

Redner, S. (1998). How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, *4*(2), 131–134.

Redner, S. (2004). Citation statistics from more than a century of physical review. *arXiv preprint physics/0407137*.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, *42*(3/4), 425–440.

Sinatra, R., Deville, P., Szell, M., Wang, D., & Barabási, A.-L. (2015). A century of physics. *Nature Physics*, *11*(10), 791.

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).

Thelwall, M., & Wilson, P. (2014). Distributions for cited articles from individual subjects and years. *Journal of Informetrics*, *8*(4), 824–839.

Vazquez, A. (2001). Statistics of citation networks. *arXiv preprint cond-mat/0105031*.

Vieira, E. S., & Gomes, J. A. (2010). Citations to scientific articles: Its distribution and dependence on the article features. *Journal of Informetrics*, *4*(1), 1–13.

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 Applications*, *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.

Yan, E., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the Association for Information Science and Technology*, *63*(7), 1313–1326.