Journal of Informetrics 8 (2014) 650–658

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Generalized preferential attachment considering aging

Yan Wu[a], Tom Z.J. Fu[b], Dah Ming Chiu[a,*]

[a] Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong
[b] Advanced Digital Sciences Center, Illinois at Singapore Pte. Ltd., Singapore

ARTICLE INFO

Article history:
Received 7 August 2013
Received in revised form 3 June 2014
Accepted 3 June 2014

Keywords:
Preferential attachment
Aging effect
Citation network

ABSTRACT

Preferential Attachment (PA) models the scientific citation process. In the PA model, a new paper attaches itself to the citation network based only on the popularity of the currently existing papers. This invariably leads to a network whose degree distribution satisfies the Power Law. Yet, empirical results show that paper age should also play a role in the citation process. In other words, when references are chosen for a new paper, the age of an existing paper may also affect the choice for citing. In this paper, we derive a generalized PA model that includes the effect of aging, with analytical solution. Such a model can be used to analyze the competing influence of preferential attachment and aging effect quantitatively in citation process and explain differences in various research domains by the extent of aging. It may also serve as a general model of network formation.

1. Introduction

In a citation network, each paper is a vertex in a directed graph, and references of each paper are the edges in the graph. When one paper cites another, an edge will be established from the referring paper to the referred one. With new papers appearing constantly, the citation network grows over time. Among different models to describe the growth process, the PA model (Barabási & Albert, 1999) is a widely accepted one. It singles out preferential attachment as the most salient feature in the citation process of scientific publications. When a paper is attached to (cited) by many others, it starts to build a cumulative advantage over those with fewer attachments (de Solla Price, 1965), which is sometimes referred to as the "rich-get-richer" property. The success of the PA model is established by correctly predicting an important property of a citation network – the degree distribution satisfies the Power Law. However, the PA model seems to be over-simplified by ignoring other factors influencing authors' citing behaviors. Actually, when an author writes papers, his choice of references is far from random (Ding, Liu, Guo, & Cronin, 2013). Factors that influence an author's choice to cite a paper may include age of the paper, research field, journal of publication, etc. The decision could also be article-dependent or reader-dependent (Bornmann & Daniel, 2008). Bethard and Jurafsky (2010) analyzed the features affecting authors' citing behaviors and found that authors tend to cite articles with more citation-counts, smaller age, and articles they have cited before. White and Wang (1997) pointed out that the cited documents are on the whole more recent.

Among the various factors, we identified an important one in this paper – aging. The PA model assumes that the probability to attach to a vertex (paper) is proportional to the vertex's degree, independent of the vertex's age (i.e. the time passed since the publication of a paper). But many researchers find evidence that the aging effect may play a role in the process of network expansion and affect the vertex degree distribution (Chiu & Fu, 2010; Dorogovtsev & Mendes, 2000; Eom & Fortunato, 2011;

* Corresponding author at: Department of Information Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong. Tel.: +852 39438357.
E-mail addresses: wy011@ie.cuhk.edu.hk (Y. Wu), fuzhengjia@gmail.com (T.Z.J. Fu), dmchiu@ie.cuhk.edu.hk (D.M. Chiu).

http://dx.doi.org/10.1016/j.joi.2014.06.002
1751-1577/

Hajra & Sen, 2005, 2006; Newman, 2009; Peterson, Pressé, & Dill, 2010; Pollmann, 2000; Redner, 2005; Wang, Yu, & Yu, 2008, 2009). For example, in Hajra and Sen (2005), the authors studied the references of various papers and found that a new paper prefers to cite a paper published in the most recent 10 years rather than a paper published earlier; the authors in Eom and Fortunato (2011), Peterson et al. (2010), Pollmann (2000), Wang et al. (2008) pointed out the pattern of how papers are forgotten over time based on experimental data.

Thus we start our paper with the hypothesis that as a paper "ages", while its ability to attract new citations still depends on its relative in-degree, it also gradually loses its luster relative to the newer counterparts. Although the preferential attachment mechanism still persists, its effect weakens over time. The quantitative analysis of such a competing process is addressed in this paper.

By generalizing the preferential attachment mechanism with a time coefficient, we construct a simple analytical model that allows us to compute the vertex degree distribution with the effect of aging. Based on this model, we can represent the competing influence of the preferential attachment mechanism and vertex aging effect for the citation process and illustrate analytically the extent of aging in different research fields. We also apply our model to a real publication dataset and find that our model fits the data well in many cases.

The rest of the paper is organized as follows. We first give a brief review of related work in Section 2. Then we propose our model and derive analytical solutions in Section 3. Experiment results are shown in Section 4. Finally, we conclude the paper with discussion of our results and future directions in Section 5.

2. Literature review

We now briefly review existing work on node aging effect in citation networks. Aging has been considered as an influencing factor during network evolution in many works. For example, Amaral, Scala, Barthélémy, and Stanley (2000) assumed that there is a constant probability that vertices may become inactive (i.e. vertices cannot receive new links) during each time step. Then the time a vertex may remain active decays exponentially. Chiu and Fu (2010) studied the change of total citation count $c(t)$ with year t in Computer Networking area, and found it is affected by x_t, the number of papers published in year t, and $\alpha(n)$, a probability distribution where n is the number of years separating the citing paper and receiving paper. Geng and Wang (2012) studied random aging effect on network clustering.

Since preferential attachment is a widely accepted model in citation network, there are also works studying node aging effect based on PA model, which are more relevant to our work. The original PA model focused on explaining the power law distribution of node degree, so it considered only one factor – node degree. Various subsequent works started to consider the time dimension as an additional factor. Newman (2009) studied this topic and the "first-mover advantage" is demonstrated via theoretical analysis and empirical evidence. He concluded that based on the original PA model, the most highly cited papers in a field are often those that were published first, rather than necessarily the best. This work does not consider what if the attachment probability is dependent on the age of the paper.

Some other works explicitly considered adding the aging effect to the attachment process. Generally they can be divided into two types: theoretical models considering preferential attachment with specific aging function and experimental works focused on empirical validation of the way in which aging effect is expressed.

In theoretical models considering preferential attachment with specific aging function, mainly two forms of aging effect are considered in theoretical analyses: exponential form and power law form. For example, in Dorogovtsev and Mendes (2000), the age damping function $s^{-\alpha}$ is introduced to the original PA model, with s denoting vertex age and α controlling the extent of age damping, while in Zhu, Wang, and Zhu (2003), the effect of aging is represented by an exponential decay factor, $e^{-\beta\tau}$, with τ denoting vertex age and β controlling the extent of aging. However, Dorogovtsev and Mendes (2000) focused most of their work on theoretical aspect, with no validation of their assumptions in real world dataset. Although Zhu et al. (2003) tested their aging function in the ISI dataset, they did not provide a closed-form solution from which the competition process of preferential attachment and node aging effect can be seen clearly.

In order to better represent the aging effect, many experiments on real-world dataset are also conducted, which are focused on empirical validation of aging effect forms. Most of those works are in favor of the exponential decay form in citation network. For example, Zhu et al. (2003) showed the exponential decay form of node aging by curve fitting of citation rate data of the years 1987–1998 obtained from the ISI dataset. Hajra and Sen (2006) studied the age distribution of references made by a paper $T(t)$ and made to a paper $R(t)$, and find by numerical simulation that an exponentially decaying $T(t)$ would give a desired $R(t)$, which should be power law decay. Wang et al. (2008, 2009) pointed out the attachment rate $\Pi(k, t) \propto k * exp(-\lambda t)$ from empirical evidence in research journals of different fields. Martin, Ball, Karrer, and Newman (2013) studied the citation patterns in the Physical Review family and also found that the citation rate drops off approximately exponentially over time.

Considering the general complex network, there are also works studying other factors influencing network structure during the evolution of networks, such as Dorogovtsev and Mendes (2001) on the accelerating growth of networks, Bianconi (2005), Dorogovtsev and Mendes (2004) on the effect of weighted network edges and Egghe (2005, 2007) on the evolution of information production processes. Since the more general complex network case is out of the scope of our current model, we do not go into details of them.