Preferential attachment in the citation network of scientific articles

N G Scherbakova and S V Bredikhin

1 Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia

E-mail: scherbakova@sscc.ru

Abstract. We study the citation network of scientific articles in the hypothesis that the likelihood of connecting to a node depends on the node’s degree, a mechanism called the preferential attachment. We show that the rate at which articles acquire citations linearly depends on the number of citations already received and we find the functional form of the dependence basing on the analysis of the dynamic source data. We also investigate the effect of the article age on receiving citations and show that there is the exponential decay of interest to old articles. Nevertheless the distribution of citations has the power-law form.

1. Background

The subject of our consideration is the citation network of scientific articles based on the data provided by the bibliographic database. The nodes correspond to scientific articles, and directed links to citations. Citation networks are characterized by a large number of nodes and a low density of links. The indegree distribution of such network is different from the distribution typical for random networks namely follows a power law

$$P(k) \sim k^{-\gamma}$$

for large $k$ ($k > k_{\text{min}}$), so networks are scale-free. The indegree of a node is the number of citations to a publication. Our aim is to examine the time evolution of citations and to investigate at which rate articles acquire new citations.

Simon [1] assumed that the principle of "having much gets more" leads to a scale-free structure of a network. Price [2] studied this mechanism as applied to citation networks and called the strategy by which success breeds success, cumulative advantage. He formulated the strategy as follows: the speed with which articles receive new citations is proportional to the citations already received. As a result the fraction of articles that have received $k$ citations (for sufficiently large $k$) decreases proportionally to $k^{-\gamma}$, where $2 \leq \gamma \leq 3$. Papers [3–5] present the results of the analysis of real citation networks that display power-law distributions of node degrees.

2. Scale-free model

Thanks to a series of works by Barabási and Albert the mechanism of cumulative advantage was called the preferred attachment, hereafter PA. In their seminal work [6] the authors considered networks as evolving dynamical systems that grow by preferential attachment. These ideas inspired the introduction of the scale-free (SF) model that has a power-law degree distribution. It is assumed that initially there are $m_0$ nodes that are connected at random, each node has at least one link. The
algorithm of the SF model is based on two principles. (1) Growth: at every timestamp a new node with \( m \) \( (m \leq m_0) \) edges is added to the network, the edges link this node to \( m \) different existing nodes. (2) Preferential attachment: the probability \( \Pi \) that a new node will be connected to node \( i \) depends on the degree \( k_i \) of node \( i \) as

\[
\Pi(k_i) = \frac{k_i}{\sum_j k_j}.
\]

So the rate with which a node with \( k \) links acquires new links is a monotonically increasing function of \( k \), this is linear preferential attachment. According to the SF model, after \( t \) timesteps, the network will have \( n = t + m_0 \) nodes and \( m_0 + (mt) \) edges. Assuming that \( k_i \) is a continuous real variable (the continuum approach), the evolution of the degree \( k_i \) of a node \( i \) can be represented by the equation

\[
\frac{d}{dt}k_i = m\Pi(k_i).
\]

With using the continuum method the authors showed that the network evolves into a scale invariant state with the probability that a node has \( k \) links following the power law with the exponent \( \gamma = 3 \), moreover \( \gamma \) is the only parameter of the model.

The generalized PA, nonlinear preferential attachment, was presented in [7] as

\[
\Pi(k_i) = \frac{k_i^\alpha}{\sum_j k_j^\alpha} = C(t)k_i^\alpha,
\]

where \( C(t) \) is a normalization constant. The analytical study [7] demonstrate that the scale-free nature of the network is destroyed for nonlinear preferential attachment. The only case in which the topology of the network is scale-free is when the preferential attachment is asymptotically linear.

2.1. Aging

Dorogovtsev and Mendes [8] proposed that for some real networks the probability that a new node connects to a node \( i \) also depends on an age of node \( i \), for example publications gradually lose their ability to attract the attention. So

\[
\Pi(k_i,t) \propto k_if(t_i),
\]

where \( f(t_i) \) is an aging function, \( t_i \) is a timestep a node \( i \) was added to the network. In the network model [8] the probability \( \Pi \) decays as \( (t - t_i)^\nu \), where \( t \) — a timestep a new node is added, \( \nu \) is a tunable parameter. The calculations predict that the degree distribution depends on the exponent \( \nu \): a power-law scaling presents only for \( \nu < 1 \), and its exponent \( \gamma \) depends on \( \nu \). Moreover for \( \nu > 1 \) the power-law scaling completely disappears, the degree distribution approaches an exponential. The exponential form of an aging function is considered in [9, 10].

2.2. Initial attractiveness

In real networks there is a nonzero probability that a new node attaches to an isolated node \( i \), so \( \Pi(0) \neq 0 \). Speaking of a citation network a node that has no citations can receive some in future. Thus [11] presented the model of a directed network such that PA has a shifted linear form:

\[
\Pi(k_i^{in}) = \frac{k_i^{in} + k_0}{\sum_j(k_j^{in} + k_0)}.
\]

where \( k_i^{in} \) is the indegree of the node \( i \), \( k_0 \) is the initial attractiveness of the node \( i \). It was shown in [11] that the indegree distribution follows a power law, and an exponent \( \gamma \) depends on \( k_0 \).

So for a model of a directed network with shifted preferential attachment and aging the probability that a newly added node will connect a node \( i \) providing that \( t \) timesteps have passed after adding \( i \) may be proportional to the following characteristics:

\[
\Pi(k_i^{in},t) \propto (k_i^{in} + k_0)f_i(t).
\]
3. Empirical research

We study quantitative features of the real citation network of articles based on data provided by RePEc database [12]. For this citation network we know the time at which each node joined the network. Our main task is to determine how the factors presented in (7) affect the network evolution. Note that these factors are not all that affect the receiving of citations, as an example: citation norms, validity, personal preferences, etc.

Hereafter we use the following terminology. If the article in its reference list contains the reference (cites) to some article, then such reference is called the outgoing citation from the point of view of the citing article, and the incoming citation from the point of view of the cited article. Let the article \( j \) cites the article \( i \). Let \( t_j \) and \( t_i \) be the years of publication of articles \( j \) and \( i \), respectively. The age of a citation \( \tau_i \) is defined as the difference in the publication year of the citing paper and the publication year of the cited paper: \( \tau_i = t_j - t_i \).

3.1. General information

The investigated data covers the period 1874 through June 2019. Total number of articles is 1404431. The data was corrected: articles with unobvious year of publication, links to the wrong year and self-citations were ignored. So we study the maximal connected component \( N_{\text{REP}} = (V, E) \), the number of nodes \( |V| = 819207 \) and the number of edges \( |E| = 5538043 \). Only internal citations are taken into account, i.e. citations between the articles contained in \( V \). The number of articles without citations is 191468 (23.37%). Number of articles cited at least once – 627739. The average number of incoming citations is 6.76. For articles cited at least once – 8.82. The average citation age \( \langle \tau_i \rangle = 11.02 \). Most cited articles have a citation age significantly higher than average. For articles with more than 500 citations, the average citation age \( \langle \tau_i \rangle = 22.87 \).

\( N_{\text{REP}} \) is characterized by an exponential growth in the number of articles and, accordingly, new outgoing citations with time. For example, in 2000 the database was replenished with 15445 articles and 57086 outgoing citations (3.69 references per one article on average), and in 2010, 35406 articles and 249324 new citations (7.04 references). The number of incoming citations to articles published during these years has increased: 175578 and 242655, respectively. When considering the distribution of incoming and outgoing citations by years from 1937 to 2019, it turned out that the curves have similar growth trends. The close correlation of distributions shows that recently published articles are cited more often than older ones. For articles published in selected years with the integrated citation distribution we see that all of them are similar except for the large \( k \) values.

We show that the incoming citation distribution for the set \( V \) follows the power law. The method for recognizing the presence of a power law in an empirical data and finding parameter values was proposed in [13]. Using the igraph package [14], it was found that the indegree distribution follows the power law with parameters \( \gamma = 2.89 \), \( k_{\text{min}} = 207 \). So the power law fits the articles having a number of incoming citations greater than or equal to 207. When we compare the citation distribution of articles published in selected years with the integrated citation distribution we see that all of them are similar except for the large \( k \) values.

3.2. Measuring \( \Pi(k) \)

Let us study (7) independently of \( k_0 \) and \( f(t) \) and prove that \( \Pi(k^w) \) is the monotonically increasing function of \( k^w \) (hereinafter we use notation \( k \) instead of \( k^w \)). In order to extract the functional form \( \Pi(k) \) from the dynamic data we use the methods presented in [15, 16]. Taking into account (3), (4), the probability \( \Pi(k) \) is considered as the speed \( A_k \) with which the existing node with the degree \( k \) acquires new edges (citations) in the process of the network growth. So we want to show that for the network under consideration \( A_k \propto k \).

We study the attachment of new nodes within relatively short time-frames in order to ignore the normalization constant \( C(t) \) that depends on the time at which a given node joins the network. Let \( T_1 = 2018 \), \( \Delta T = 1 \) and consider outgoing citations of articles published between \( [T_1, T_1 + \Delta T] \), that is, during 2018. The number of outgoing citations for this period is 658415. Let \( T_0 = T_1 - 1 \), i.e. \( T_0 = 2017 \). For each article published within time window \( w_1 = [1874, T_0] \), we calculate the number of
incoming citations \( k \) received during \( T_1 \). Then we find \( \Delta k \) – the average number of incoming citations received by articles with \( k \) citations within the period \([ T_1, T_1 + \Delta T]\). We consider additional time windows \( w_2 = [1988, T_0], w_3 = [1998, T_0], w_4 = [2008, T_0] \), that is, we change the initial year of the period for which \( k \) is calculated and then calculate \( \Delta k \). The results of \( A_k \) approximation with a linear function with use the method of least squares show that \( A_k \) has a form \( a_k + b \), see Tab. 1, column \( f_1 \). It should be noted that the approximating function depends on the time window. Moreover, it depends on the range of \( k \), see Tab. 1, column \( f_2 \) for \( k \leq 50 \). We can present the linear form \( a_k + b \) as \( k + b/a \), that is, \( A_k \sim k + k_0 \), where \( k_0 \) can be considered as the initial attractiveness. It makes sense for small values of \( k \) \((k \leq 50)\), since in this case \( k_0 \) is positive.

### Table 1. Summary of the approximating functions for \( A_k \). It can be seen that the functional form depends on the time window \((w_1 - w_4)\) and the range of \( k \) (all values or \( k \leq 50 \))

| Window | \( f_1 \) | \( f_2 \) |
|--------|----------|----------|
| \( w_1 \) | \( y = 0.090066x - 0.404866 \) | \( y = 0.094525x + 0.369329 \) |
| \( w_2 \) | \( y = 0.090105x - 0.405038 \) | \( y = 0.094847x + 0.365893 \) |
| \( w_3 \) | \( y = 0.095328x - 0.959371 \) | \( y = 0.098202x + 0.345890 \) |
| \( w_4 \) | \( y = 0.126907x - 0.682570 \) | \( y = 0.127795x + 0.213792 \) |

We can see that the estimation is unstable. The choice of several time windows for calculating \( k \) allows to give an aggregated estimation, but in this case \( C(t) \) value changes will be ignored.

### 3.3. Aging function

We study the aging (obsolescence) process, ignoring the PA mechanism and calculate the period of time articles with the same number of citations will be cited again. An analysis of the relationship between the average citation age and the number of citations received reveals a positive correlation, which can be clearly seen for the number of citations less than 50 (approximating function \( y = 0.08882x + 7.3354 \)). For large values, significant fluctuations are observed, and the amplitude of fluctuations increases. “Outstanding” articles (often cited) have a large average citation age. That is, the processes of articles aging and obsolescence are different.

We study two distributions: \( S(t) \) and \( D(t) \). \( S(t) \) is the distribution of ages of cited articles that receive citations from articles published in the selected year. \( D(t) \) is the distribution of ages of citing articles that cite articles published in the selected year.

Let us fix the year of publication of citing articles \( t_0 \). Articles published in the year \( t_0 \) cite \( n \) articles published in years \( t_1, t_2, ..., t_j, ... \) (in the past with respect to \( t_0 \). Let \( n_i \) be the number of cited articles published in the year \( t_i \). The distribution of intervals \( \theta_i = (t_i - t) \) defines \( S(t) \). Trends of forgetting past articles can be seen in the presented figure. Abscissa indicates the intervals that have passed since the

![Figure 1](image_url)

**Figure 1.** The distribution \( S(t) \) of the ages of past articles cited by papers published in the selected years. It can be seen that the curves related to different years are similar.
publication of cited articles, and ordinate shows the proportion of cited articles \( n_i/n \), the logarithmic scale. The figure shows that citing articles prefer to cite papers published 2–3 years before, then the exponential decay follows. The distribution \( S(t) \) can be considered as an analogue of an aging function (5), i.e., the aging function has the form \( f(t) = \exp(\alpha t) \), \( \alpha < 0 \). It was shown [8] that for a network with a constant growth rate, an exponential aging function leads to an exponential distribution of the degrees of nodes. However, it was found that for \( N_{REP} \) the indegree distribution follows the power law. As noted in [16] due to the exponential growth in the number of publications with time, the network is less affected by the rapid dropping in interest to past articles.

Now let \( t_0 \) be the year of publication of articles for which incoming citations are considered, these are cited by articles published in years \( t_1, t_2, \ldots, t_n, \ldots \) (in the future, with respect to \( t_0 \)). Let \( n_i \) be the number of citing articles published in the year \( t_i \). The distribution of the intervals \( \theta_i = (t_i - t_0) \) defines \( D(\theta) \). The analysis of the distribution shows that during the first 21 years there is the linear increase, followed by an exponential decay, but the decay is much slower than for \( S(t) \). That is, citing authors forget about publications faster than it affects cited publications. As a result, the indegree distribution follows the power law.

**Summary**

The computational experiment is devoted to the analysis of processes that influence the evolving of the citation network of articles \( N_{REP} \) based on RePEc data. We focus on the PA mechanism and find that the growth of citations to an article is proportional to its current number of citations and is approximated by the linear function.

On the one hand, according to the idealized PA mechanism, articles with the high number of citation are most likely to receive new ones, on the other hand the exponential increase in the number of articles reduces the chance of citing old articles, and the main increase in the number of citations occurs in the first two years after publication. For \( N_{REP} \) the average citation age is about 11 years. The aging process can be represented by the decaying exponential function, exponent values depend on the range of ages of the cited articles. It should be noted that some articles are currently being cited at a significant rate, and some articles are “dead”. If we consider citing ages to articles published in the certain year we see the exponential decay 21 years after publication of cited articles, but the decay is slower. So in spite of exponentially-decaying interest to past publications, due to exponential growth of publications and citations, the number of citations to past publications is increasing.

**References**

[1] Simon H A 1955 On a class of skew distribution functions *Biometrika* 42 425–40
[2] Price D J 1976 A general theory of bibliometric and other cumulative advantage processes *J. Amer. Society for Inform. Sci.* 27 292–306
[3] Tsallis C and de Albuquerque M P 2000 Are citations of scientific papers a case of nonextensivity? *Eur. Phys. J. J.* 13(4) 777–80
[4] Redner S 1998 How popular is your paper? An empirical study of the citation distribution *Eur. Phys. J. B* 4(2) 131–134
[5] Peterson G J, Presse S and Dill K A 2010 Nonuniversall power law scaling in the probability distribution of scientific citations in *Proc. Natl. Acad. Sci. USA* 107 (37) 16023–27
[6] Barabási A-L and Albert R 1999 Emergence of scaling in random networks *Science* 286 509–12
[7] Krapivsky P L, Redner S and Leyvraz F 2000 Connectivity of growing random networks *Phys. Rev. Lett.* 85 4629–32
[8] Dorogovtsev S N and Mendes J F F 2000 Evolution of reference networks with aging *Phys. Rev. E* 62. 1842–45
[9] Zhu H, Wang X, and Zhu J-Y 2003 Effect of aging on network structure *Phys. Rev. E* 68 056121
[10] Hajira K B and Sen P 2006 Modelling aging characteristics in citation networks *Physica A* 368 575–82
[11] Dorogovtsev S N, Mendes J F F and Samukhin A N 2000 Structure of growing network with preferential linking Phys. Rev. Lett. 85, 4633–36
[12] RePEc http://repec.org
[13] Clauset A, Shalizi C R and Newman M E J 2009 Power-law distributions in empirical data SIAM Review 51 661–703
[14] igraph – The network analysis package https://igraph.org
[15] Jeong H, Neda Z and Barabási A L 2003 Measuring preferential attachment for evolving networks EuroPhysics Letters 61 567–72
[16] Redner S 2004 Citation statistics from more than the century of Physical Review arXiv:physics/0407137