Computer Physics Communications 182 (2011) 185–187

Contents lists available at ScienceDirect

Computer Physics Communications

www.elsevier.com/locate/cpc

Dynamics of citation distribution

Jiann-wien Hsu [a], Ding-wei Huang [b,*]

[a] General Education Center, National Tainan Institute of Nursing, Tai-nan, Taiwan
[b] Department of Physics, Chung Yuan Christian University, Chung-li, Taiwan

ARTICLE INFO

Article history:
Received 14 January 2010
Received in revised form 10 June 2010
Accepted 16 July 2010
Available online 29 July 2010

Keywords:
Citation analysis
Bibliometrics
Stochastic processes

ABSTRACT

We study the citation dynamics of scientific publications over the years. We propose a simple cellular automaton model featuring a combination of two distinct mechanisms, i.e. the random assignment and the preferential attachment, to investigate the dynamics of journal citation. Different from most previous studies focusing on highly cited papers, we analyze the time evolution of the entire citation distribution. Empirical data can be well reproduced by numerical simulations. Within the linear regime of the Cited Half-Life, a steady accumulation of citations can be expected. Moreover, within this linear regime, the ratio between the above two mechanisms is a constant. Besides the average citation represented by the Impact Factor, such a constant ratio can also be a characteristic of the journal.

1. Introduction

Publishing is serious business to researchers in these days. With so many researchers working hard to advance our knowledge, the amount of publications increases rapidly. There are many emerging patterns in such a human academic activity [1,2]. Besides the social science, it is not surprising that scientific publishing also becomes an interesting topic in statistical physics [3–6]. One of the correlations among scientific publications is through the citation. Researchers share credit by citing each other's works in their papers. Obviously, every researcher wants his/her works to be cited numerously. Citation number therefore plays the role as an indicator to the impact of a published article, presumably reflecting the importance of a research work. Consider all publications from the scientific community, it is easy to expect a wide range of citation numbers [7]. The so-called well-known papers accumulate a large number of citations over the years; while some unknown papers might never be cited by others. Such a wide spectrum can be expected to contain various professional journals. With naive expectation, each journal occupies a much narrower range of the above spectrum. And the prestige of a journal can be recognized by the citation number to articles published in that journal. Then, it is interesting to ask how the citation distribution may still have a wide distribution for articles published in the same journal, which supposes to review and publish submitted articles on the basis of equal criteria.

To our knowledge, the citation statistics was first studied more than forty years ago by Derek J. de Solla Price [8], whose work has indicated a power-law distribution. Later, the author proposed the so-called Cumulative Advantage Processes to understand the dynamics of citation [9], where a statement of "a paper which has been cited many times is more likely to be cited again than one which has been little cited" was presented. The mechanism could also be understood as preferential attachment in the framework of evolving networks [10]. More recently, Laherrère and Sornette [11] presented evidence for a stretched exponential distribution. Redner [12] suggested a power-law decay for the large citation tail. Tsallis and de Albuquerque [13] proposed a continuous distribution from the non-extensive thermostatistical formalism. The empirical data studied in those works either included a wide range of journals or accumulated over a long period of time. Most previous studies focused on the highly cited papers. The portions of small citations were often neglected. Some even conjectured that those papers with zero citation should be separated individually [2]. In this work, we propose a simple cellular automaton model combining two distinct mechanisms, i.e. the random assignment and the preferential attachment, to address the entire citation distribution.

2. Model

Consider an ensemble of N articles, which receives a total number of M citations. Let the average number of citations for each article be $a = M/N$. We assume that these N articles cover a wide range of research topics. We also assume that these N articles are equal in the quality and thus have the same potential to be cited by other research works. Here we discuss two distinct mechanisms for the dynamics of citation. If these M citations are randomly assigned, the resultant distribution should be a binomial distribution. Such a citation distribution has a narrow width with its peak located at the average number a. As the citations accumulate, the peak of distribution shifts accordingly. The mechanism of random

[*] Corresponding author.
E-mail address: dwhuang@phys.cycu.edu.tw (D. Huang).

0010-4655/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.cpc.2010.07.031

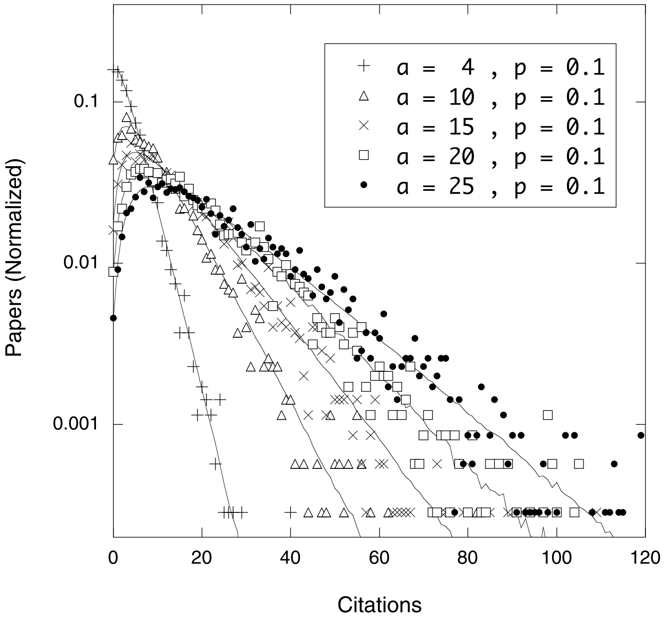

Fig. 1. Results of the simulations with $p = 10\%$. The solid lines show the result of averaging 1000 simulations.

citations is plausible, to which we do not imply that the literature references at the end of each research article are randomly assembled [14]. In practice, we believe that every author arranges the references deliberately. However, as the research interests and topics are different from article to article, the citations therein might look like randomly assigned. Especially these N articles are presumed to have the same quality of research and the same level of impact to other researchers, which is often measured by the number of citations. Thus it can be quite reasonable to expect that each of these articles will receive more or less equivalent number of citations, which results in the narrowly-spread binomial distribution.

Next, we consider a different mechanism for the dynamics of citation. When these articles first appear in publication, they are expected to be cited equally. However, once an article is cited, it will be further cited more easily. That article can now be referred to not only through the original publication but also through those works having cited it as a reference. With this mechanism of preferential attachment, the resultant distribution is an exponential distribution. With this mechanism at work, the citations will shift swiftly from weighing equally in the very initial stage to focusing on those cited frequently in a later stage. If an article is not cited in the early stage, it will be even harder to receive any citations in the later stage. Thus the mechanism prescribes a monotonically decreasing distribution as the citations increase. The exponential distribution has its highest weight at the zero citation. As the citations accumulate, large citations appear swiftly.

We propose a simple cellular automaton model to combine these two plausible mechanisms. Consider N newly published articles without citation records. At each discrete timestep, one citation is given to one of these articles. At the end of M timesteps, citation distribution over these N articles will be analyzed. Each citation will be assigned by stochastically switching between two different mechanisms. In the first mechanism, all the articles are equally weighted. Each article has the same probability to receive the citation. In the second mechanism, the weighting is related to the citation history. For each article, the weighting is proportional to one plus the cumulative citation number up to the previous time step. Each article has a different weighting. The most cited article will have the highest probability to receive the citation.

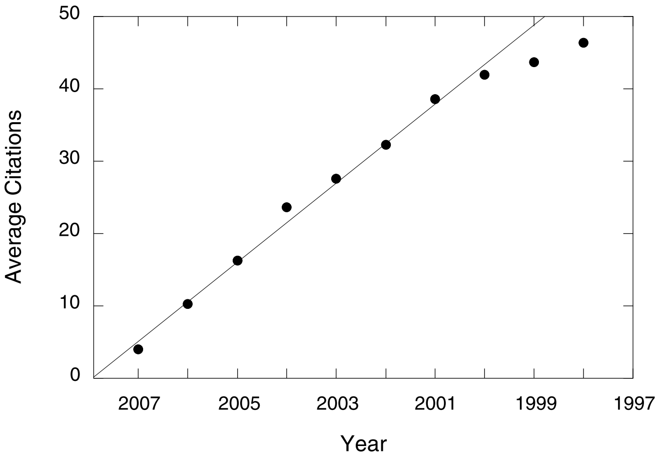

Fig. 2. Average citation number to research articles published in *Phys. Rev. Lett.* in different years. The data from 2007 to 2000 can be well described by a linear fit with a slope of 5.5 citations per year.

The mixing between these two mechanisms is controlled by the parameter p. At each time step, the first mechanism will be activated with a probability p; the second mechanism will be used with a probability $(1 - p)$ instead. The results of simulations are shown in Fig. 1, where we show both the data points in one simulation and the smooth curves averaged over 10^3 simulations. The binomial distribution and exponential distribution can be restored, respectively, in the limits $p = 1$ and $p = 0$. For large citations, the distribution is characterized by an exponential decay. For small citations, the distribution increases with the increase of citation numbers, which is a characteristic of binomial fluctuations.

3. Empirical data

We analyze the time evolution of citation distribution to published papers. The research journal *Physical Review Letters* publishes around 3500 research articles annually. As published in the same journal, each of these articles goes through the same process of being refereed and accepted. It is fair to assume that all these 3500 articles have the same potential to be cited by other research works. Up to July 2008, the average citations to these papers are shown in Fig. 2. To reveal the dynamical effects, citation counts to papers published in this journal in different years are separated. Up to the first half year of 2008, each paper published in 2007 has been cited four times on average. The average citation number for papers published in 2006 increases to ten. For earlier published papers, the average citations increase accordingly. From 2007 to 2000, the data can be fairly described by a linear relation with a slope of 5.5. For papers published much earlier, i.e. before 1999, deviations to the linear increase of accumulated citations can be noticed. As an ergodic distribution, such accumulated citations for different papers published in preceding years can be reinterpreted as the expected citations for a newly published paper in the coming years. For a newly published paper, it can be expected that 5.5 citations will be accumulated each year for up to seven consecutive years. Beyond the linear increase, the accumulated citations are expected to slow down a bit beyond the seventh year.

The citation distributions are shown in Fig. 3, where the corresponding average citations are shown in Fig. 2. For the papers published in 2007, the citations up to the first half year of 2008 distribute as an exponential decay. However, the zero citations are overestimated by the exponential distribution. For the papers published in earlier years, the citations begin to accumulate. The distribution on the higher citations enhances with the suppression on the lower citations. The empirical data can be well described

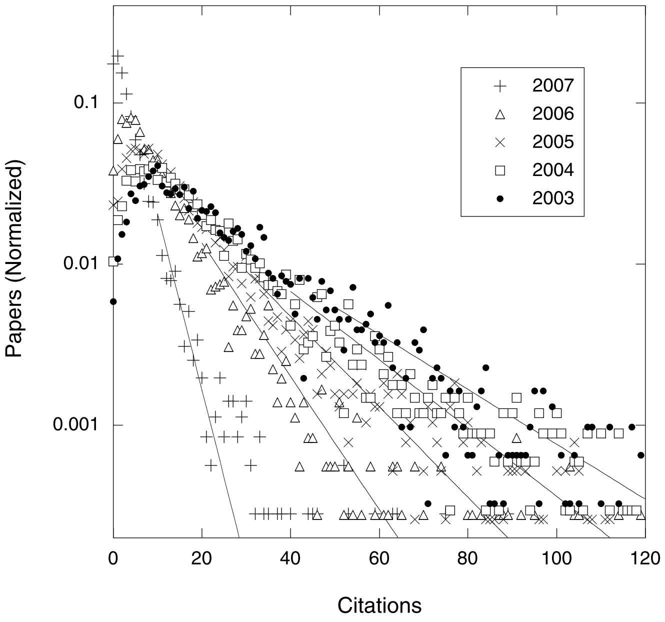

Fig. 3. Citation distributions to papers published in different years. With the corresponding average citations from Fig. 2, the solid lines show the exponential tails.

by our model as shown in Fig. 1, for both the fluctuations in a single simulation and the smooth curves by averaging many simulations. Within the linear regime, the empirical data can be fairly reproduced by a fixed ratio $p = 10\%$. By a single mechanism, either the random assignment in binomial or the preferential attachment in exponential, the empirical data cannot be reproduced. However, by mixing these two mechanisms stochastically, the entire distribution can be fairly described.

4. Discussions

In this paper, we propose a simple model to investigate the time evolution of citation distributions. From large citations to small citations, including the zero citation, the entire distribution can be fairly described in the same framework. The average number of citations characterizes the exponential tail in large citations. The distribution in the small citations is controlled by the ratio between the two mechanisms: random assignment and preferential attachment. It would be interesting to further investigate the parameterization for each dataset, which might reveal the time dependence of those parameters. A mean-field theory is under progress to control the noise from the intrinsic fluctuations in citation dynamics. In contrast to most previous studies, we do not trace the citation history of specific papers. Instead, we analyze the accumulated citations to articles published in previous years. The annual citations can be related to the well-known Impact Factor, which is a conventional measure of the citations to a journal. The Impact Factor can be taken as the average citation for the recently published papers. More specifically, the Impact Factor is defined as the number of citations in a year given to those papers published during the two preceding years. The annually accumulated 5.5 citations shown in Fig. 2 can be translated into an Impact Factor of 8.3.[1] We also notice that the linear increase of accumulated citations covers the so-called Cited Half-Life.[2] Within the Cited Half-Life, a steady accumulation of citations can be expected. We also observe that, within this linear regime, the ratio between the two mechanisms is a constant. Besides the average citation a, this ratio p can also be a characteristic of the journal.

References

[1] F. Radicchi, S. Fortunato, C. Castellano, Proc. Natl. Acad. Sci. 105 (2008) 17268.
[2] S. Redner, Physics Today 58 (2005) 49.
[3] M. Wang, G. Yu, D. Yu, Physica A 387 (2008) 4692.
[4] K.B. Hajra, P. Sen, Physica A 368 (2006) 575.
[5] P.L. Krapivsky, S. Redner, Phys. Rev. E 71 (2005) 036118.
[6] S. Lehmann, B. Lautrup, A.D. Jackson, Phys. Rev. E 68 (2003) 026113.
[7] W. Shockley, Proc. IRE 45 (1957) 279.
[8] D.J. de Solla Price, Science 149 (1965) 510.
[9] D.J. de Solla Price, J. Amer. Soc. Inform. Sci. 27 (1976) 292.
[10] A.L. Barabási, R. Albert, Science 286 (1999) 509.
[11] J. Laherrère, D. Sornette, Eur. Phys. J. B 2 (1998) 525.
[12] S. Redner, Eur. Phys. J. B 4 (1998) 131.
[13] C. Tsallis, M.P. de Albuquerque, Eur. Phys. J. B 13 (2000) 777.
[14] M.V. Simkin, V.P. Roychowdhury, Ann. Improb. Res. 11 (2005) 24.

[1] The Impact Factor of *Phys. Rev. Lett.* is 6.944 in 2007. In this study, we consider only the citations to research articles. While the ISI (Institute for Scientific Information) database also includes other types of documents, which results in the difference in average citation number.

[2] The Cited Half-Life is defined as the median age of those journal articles cited in a given year. In 2007, the Cited Half-Life of *Phys. Rev. Lett.* is 7.0 years, i.e. half of the citations to *Phys. Rev. Lett.* in 2007 are to articles published within 7 years.