Citation dynamics of scientific papers and Bass model for information diffusion

Michael Golosovsky and Sorin Solomon
The Racah institute of Physics, The Hebrew University of Jerusalem, Israel

Abstract

Our aim is to apply Bass model for diffusion of innovations to account for dynamics of citations to scientific papers. To this end we traced citation history of individual scientific papers, considered first and second generation of their citing papers, and measured the number of direct and indirect citations (innovators and imitators in the parlance of Bass model). We found that citation dynamics can be described by the Bass model with time-dependent, and most important- nonlinear coefficients. This nonlinearity brings a new and crucial element - an autocatalytic loop that can produce the runaway phenomenon- some papers are being cited practically forever. Our results can serve for forecasting the future citation behavior of a paper, group of papers, or of a journal impact factor.

The first-principle models of citation dynamics are usually based on the redirection/copying mechanism [1,2,3]. The scientist that composes the reference list of his new paper finds up some papers almost randomly (direct citations) and then picks up some of their references (indirect citations). The assumptions at the core of these models have not been verified and the value of relevant parameters were not studied yet. The verification of such grotesque scenario comes mostly from the comparison of the measured citation distributions to model prediction. However, citation distributions are not very sensitive to details of the microscopic mechanism. Here, we undertook the task to perform microscopic measurements aimed to validate the redirection mechanism and to estimate the relevant parameters. Our crucial methodological innovation is that we formulate the citation model from the perspective of a cited paper rather than from the perspective of an author. Our resulting model reduces to the Bass model for diffusion of innovations [4].

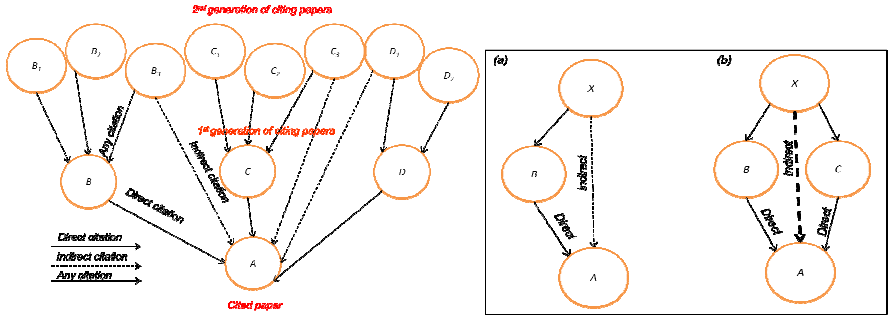

Fig.1. Left panel: Schematic representation of the model. The papers that cite paper A directly (B, C, D) appear when the citing author finds the paper A by reading the journal, digging in the databases, or following the scientific media. The paper A is cited indirectly when the citing author reads some recent papers (B, C, D) and picks up the paper A from their reference list. Right panel: A possible source of nonlinearity- the probability of an indirect citation in case (b) is more than twice higher than in case (a).

We assume that the number of citations of a certain paper during short time interval Δt follows the stochastic growth process $\Delta k=\lambda\Delta t+\sigma dW(t)$ where λ is the so-called latent citation rate [5] that obeys a differential equation

$$\lambda = p(t) + \int_0^t q(k, t-\tau)\frac{dk}{d\tau}d\tau \qquad (1)$$

where t is the number of years after publication and k is the number of citations. Here, the first and second terms account for direct and indirect citations, correspondingly. Equation (1) describes a Bellman-Harris branching process where higher-order terms corresponding to the third generation of citations have been neglected. We assume that parameters q and p depend on time. Moreover, our measurements below suggest that the parameter q depends on k (while the original Bass model assumes constant p and q).

To check the validity of Eq. 1 in the context of citation dynamics we measured the direct and indirect citations for several cohorts of similar papers that were published in the same journal, in the same year. The cohorts differ with respect to the number of citations garnered by 2013. Following the approach of Ref. [6] we considered

several generations of the citing papers. Since we are interested mostly in statistics we restricted ourselves to the first and second citation generation, identified direct and indirect citations (Fig. 1 a, b), counted them, and measured parameters p and q in Eq. 1. At the next step we considered all 40 195 Physics papers published in 1984, measured cumulative citation distribution for each year from 1984 to 2008 and compared them to model prediction based on Eq.1 and previously found parameters q and p. The model fits the data perfectly well (Fig.1c). We took actual citation distributions for first three years after publication (the publication year was considered as t=1) as initial conditions for Eq. 1.

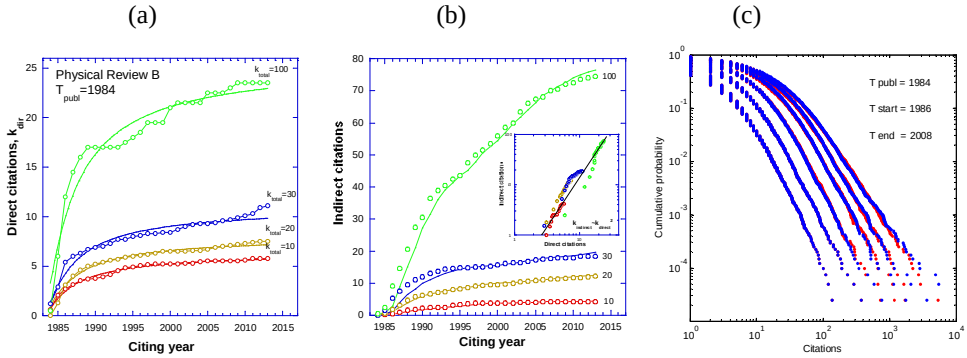

Fig.1 Citation dynamics for several ensembles of Physics papers published in 1984 in the Physical Review B (ensemble averaged). Each ensemble consists of papers that garnered the same number of citations in the period of 1984-2013 (10, 20, 30, and 100 citations). (a)- Direct citations. The continuous line shows empirical dependence $p_0 e^{-3.4/(t+0.7)}$ where p_0 is different for different cohorts. (b)- Indirect citations. Continuous lines show approximation by Eq.1 with $q=(0.2+0.12\ln k)e^{-0.8(t-1)}$. The inset shows that the number of indirect citations increases with the number of direct citations in a nonlinear way. (c) Cumulative citation distributions to 40 195 Physics papers published in 1984. Each curve corresponds to the citing year 1985, 1986, 1988, 1992, 1997, 2008. The model based on Eq.1 with the above parameters p, q (blue symbols) fits the data (red symbols) fairly well for all citing years.

Thus, our measurements validate the model given by Eq.1. The emerging citation dynamics is as follows. Immediately after publication the paper acquires mostly direct citations whose number depends on the popularity of the field, circulation number of the journal, coverage in media, reputation of the institution- in short, the factors that determine the willingness of the reader to read the paper. Each direct citation initiates the cascade of indirect citations [7]. Most of these cascades die out but due to the autocatalytic loop illustrated in the right panel of the Fig.1, the papers that already got many citations (first movers [8]) have a prolonged longevity. Some of them can achieve a tipping point in their career after which they can become practically immortal.

The similarity between information diffusion (as measured by citation dynamics) and disease propagation is well-known [6,9]. However, nonlinearity resulting from the autocatalytic loop in Eq.1 (parameter q increases with k) is something new and to the best of our knowledge does not appear in epidemiology. We speculate that similar nonlinearity can appear in other applications of Eq. 1 such as rumor propagation or viral marketing. Then, citation dynamics represents a useful playground for testing the models of such propagation.

1. P. L. Krapivsky and S. Redner, "Network growth by copying", Phys. Rev. E **71**, 036118 (2005).
2. M.V. Simkin and V.P. Roychowdhury, "A Mathematical Theory of Citing", J. Am.Soc. Inf. Sci. Techn. **58**, 1661, (2007).
3. G.J. Peterson, S. Presse, and K.A. Dill, Proc. Natl. Acad. Sci. USA **107**, 16023 (2010).
4. F. Bass "A new product growth model for consumer durables", Management Science **15,** 215 (1969).
5. M. Golosovsky and S. Solomon, Phys. Rev. Lett. **109**, 098701 (2012).
6. A. Scharnhorst, E. Garfield arXiv:1010.3525v1 (2010)
7. A. Mazloumian, Y-H. Eom, D. Helbing, S. Lozano1, S. Fortunato, PLoS ONE **6**, e18975 (2011).
8. M. E. J. Newman, "The first mover advantage" Europhys. Lett. **86**, 68001 (2009).
9. L.M.A. Bettencourt, A. Cintron-Arias, D.I. Kaiser, C. Castillo-Chavez, Physica A **364** 513 (2006).