Stochastic Dynamical Model of a Growing Citation Network Based on a Self-Exciting Point Process

Michael Golosovsky* and Sorin Solomon

The Racah Institute of Physics, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel

(Received 29 November 2011; published 28 August 2012)

We put under experimental scrutiny the preferential attachment model that is commonly accepted as a generating mechanism of the scale-free complex networks. To this end we chose a citation network of physics papers and traced the citation history of 40 195 papers published in one year. Contrary to common belief, we find that the citation dynamics of the individual papers follows the *superlinear* preferential attachment, with the exponent $\alpha = 1.25$–1.3. Moreover, we show that the citation process cannot be described as a memoryless Markov chain since there is a substantial correlation between the present and recent citation rates of a paper. Based on our findings we construct a stochastic growth model of the citation network, perform numerical simulations based on this model and achieve an excellent agreement with the measured citation distributions.

DOI: 10.1103/PhysRevLett.109.098701 PACS numbers: 89.75.Da, 02.50.Ga, 89.65.Ef, 89.75.Hc

The field of growing complex networks (informational, social, biological, etc.) has attracted increasing interest in the physics community during the past decade [1–3]. Many of these networks are believed to achieve a stationary state and to become scale-free [1,4,5]. The static characteristics of growing networks such as clustering coefficient, community structure, and degree distribution were extensively studied both theoretically and empirically [1–3] while the dynamics of these networks was studied mostly theoretically. It is widely believed that they are generated by the preferential attachment [1] (cumulative advantage [5]) mechanism. The latter assumes that new links are distributed between existing nodes with probability $\Pi_i = \lambda_i/\sum_i \lambda_i$ where λ_i is the attractivity, i.e., the expected number of links acquired by a node i in a short time interval Δt [1]. From the perspective of a single node, the number of incoming links grows according to the inhomogeneous Markov process with the rate

$$\lambda_i = A(k_i + k_0)^{\alpha}, \tag{1}$$

where k_i is the number of existing links, k_0 is the "initial attractivity," α is the attachment exponent, t is the age of the node, and $A(t)$ is the aging function [2,6]. In fact, Eq. (1) describes the stochastic multiplicative growth process

$$\Delta k_i = \lambda_i \Delta t + \sigma dW(t), \tag{2}$$

where Δk_i is the actual number of newly acquired links during time interval Δt and $\sigma dW(t)$ is its stochastic component.

The direct way to verify Eq. (1) is to measure Δk_i distributions for the sets of nodes with the same degree k, to find $\lambda = \overline{\Delta k_i}$, and to check how λ depends on k. Previous studies that were aimed at this goal [7–10], focussed on the citations to scientific papers as one of the best documented networks and a prototype for the study of dynamic behavior of growing networks [11]. Since the above studies were restricted to relatively small or inhomogeneous data sets, they had to apply indirect averaging procedures, such as numerical integration [7,8] or moving average [9,10]. These procedures are prone to quantization errors and yield inconclusive results.

Our goal is the direct measurement of the average growth rate of the node degree in a complex network [Eq. (1)] and the assessment of its stochastic part [Eq. (2)] as well. Following the accepted practice [7–10] we chose a network of citations to scientific papers. We performed a high-statistics and time-resolved study of the citation dynamics of a very large set of papers that is field and age homogeneous (one scientific discipline, one publication year). Based on our findings we constructed a stochastic model of citation dynamics with no "hidden" parameters such as fitness [12] or relevance [13]. Then we performed a numerical simulation based on our model and verified that the real and simulated citation networks have the same microscopic and macroscopic characteristics.

We used the Thomson-Reuters ISI Web of Science, chose 82 leading physics journals, excluded review articles, comments, editorials, etc., and analyzed the citation history of all 40,195 original research physics papers published in these journals in one year—1984. For each paper i we determined $k_{i,t}$—the total number of citations accumulated after t years ($t = T_{\mathrm{cit}} - T_{\mathrm{publ}} + 1$), and $\Delta k_{i,t}$—the number of citations gained by the same paper in the year $t + 1$. For every citing year t we grouped all papers into ~40 logarithmically spaced bins, each bin containing the papers with close k. Figure 1 shows statistical distributions of Δk_i for several such bins and for two selected years. For each bin we found the mean, $\lambda(k) = \overline{\Delta k_i}$, and the variance, $\sigma^2 = \overline{(\Delta k_i - \lambda)^2}$.

Figure 2 shows that $\lambda(k)$ dependence is well accounted for by Eq. (1) where A, k_0, and α are fitting parameters. We

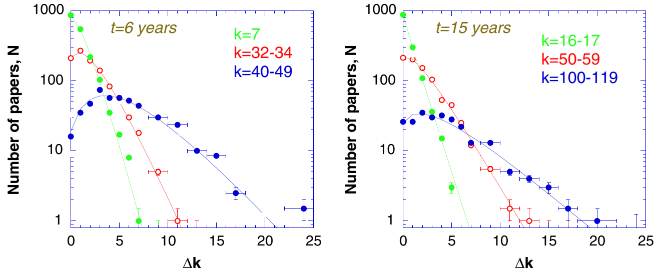

FIG. 1 (color online). Statistical distribution of additional citations Δk_i accumulated during the time window of $\Delta t = 1$ year. Continuous lines show fits to a negative binomial distribution. k is the number of previous citations and t is the number of years after publication.

found that the aging function follows the power-law decay, $A = 3.54/(t + 0.3)^2$; the initial attractivity is almost time independent, $k_0 \approx 1.1$; the exponent α gradually increases with time from $\alpha = 1$ to $\alpha = 1.25$. Although the deviation of α from unity is small, it is significant and contrasts with the assumption of linearity commonly accepted by the practitioners of the preferential attachment model [1,4,5,13,14]. Indeed, while the linear preferential attachment generates the scale-free network with the power-law degree distribution, the superlinear preferential attachment tends to generate the ''winner-takes-all'' network [2,6].

For comparison, we performed similar measurements for the mathematics and economics papers published in the same year (1984). We found that the citation dynamics for both these disciplines is also well accounted for by Eq. (1). The α and k_0 turn out to be almost the same as those for physics while the aging function $A(t)$ is different (see Supplemental Material [15]). Similar α and k_0 were found in the US patent citation studies [16]. This suggests a universal microscopic mechanism of citation accumulation whereas the variations in total citation counts between scientific fields can be attributed to different initial conditions (the number of citations gained during first couple of years after publication) and to different growth rates of the number of publications.

In what follows we analyze another key ingredient of the preferential attachment model—the Markov chain assumption. Since Eq. (1) postulates that the citation rate $\lambda = \overline{\Delta k_i}$ depends *only* on the number of previous citations k, it follows that the statistical distribution of additional citations Δk_i, gained by the papers with the same k during a time window Δt, should be nothing else but the Poissonian:

$$P(\Delta k) = e^{-\lambda \Delta t} \frac{(\lambda \Delta t)^{\Delta k}}{(\Delta k)!}. \quad (3)$$

To the best of our knowledge, statistical distribution of additional citations has not been measured so far. This new kind of measurement (Fig. 1) reveals that the Δk_i distributions are broader than the Poissonian. To quantify this broadening we used the variance-to-mean ratio, $F = \sigma^2/\lambda$, also known as the index of dispersion or Fano factor. Figure 3 shows that $F \approx 1$ for small k, as expected for the Poisson distribution, while $F \gg 1$ for large k. This strong deviation from the Poissonian indicates that Eq. (1) misses some important factor which determines the growth of citation networks. We reasoned that the missing factor is related to the citation history of papers. To probe this conjecture we considered the temporal autocorrelation of the annual citations, $\Delta k_i(t)$. Since the typical citation history of a paper is too short (10–15 years), the measurement of autocorrelation for a single paper is unreliable. Therefore, we measured autocorrelation in the sets of papers that at certain citing year t have the same number of previous citations k. Specifically, we found the number of citations garnered by each paper in such a set during the current year and the last year, $\Delta k_{i,t}$ and $\Delta k_{i,t-1}$,

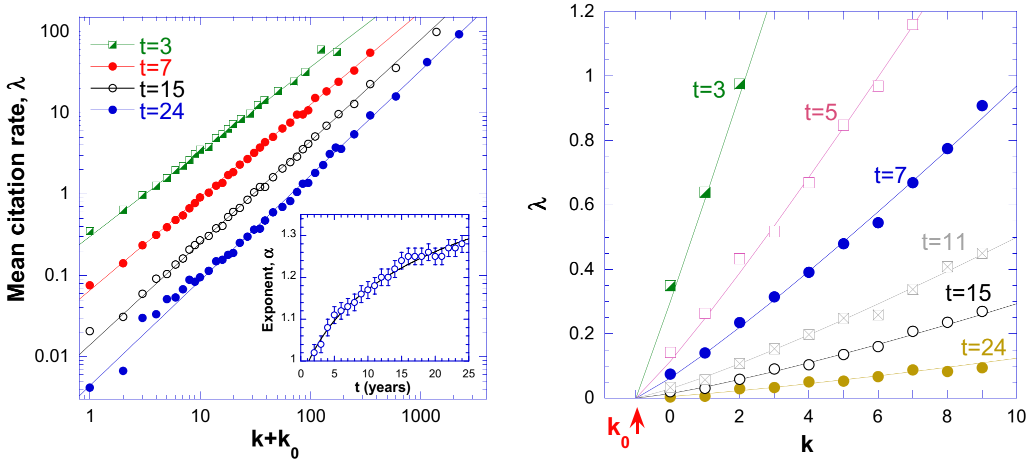

FIG. 2 (color online). Left panel: Mean annual citation rate, $\lambda(k) = \overline{\Delta k_i}$, as a function of the number of previous citations k; t is the number of years after publication. The continuous lines show a superlinear fit, $\lambda = A(k + k_0)^\alpha$ where $k_0 = 1$ and α is shown in the inset. Right panel: The same data in the linear scale. The intercept of the continuous lines with the horizontal axis yields time-independent $k_0 \approx 1$.

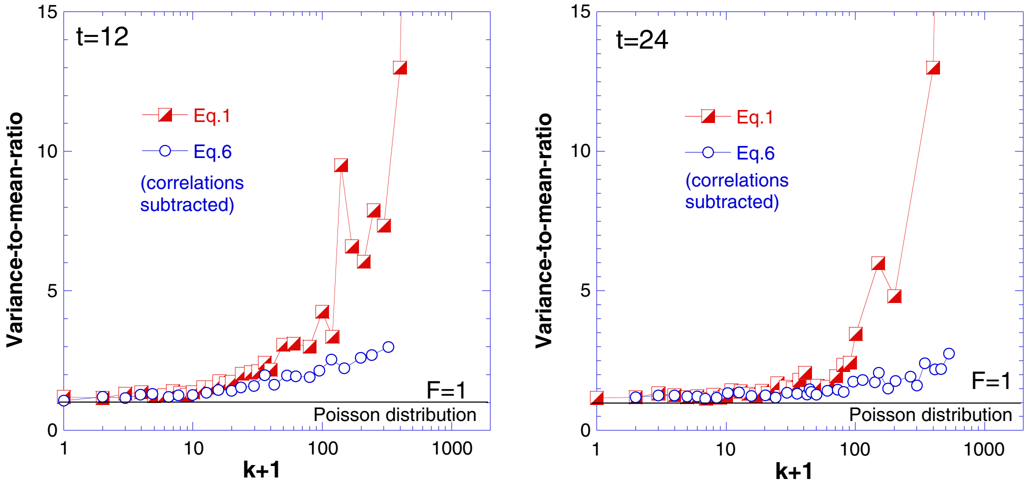

FIG. 3 (color online). The variance-to-mean ratio (Fano factor), $F = \sigma^2/\lambda$, for the statistical distributions of additional citations Δk_i (see Fig. 1). Each point corresponds to the set of papers with the same number of expected citations λ_i, given by Eq. (1) (the red squares). The data, especially for $k > 60$, deviate upwards from the $F = 1$ line, characteristic for the Poisson distribution. The blue circles show the variance-to-mean ratio for the Δk_i distributions for the sets of papers with the same number of expected citations λ_i, Eq. (6). These data are closer to the $F = 1$ line.

correspondingly, and calculated the Pearson autocorrelation coefficient

$$c_{t,t-1} = \frac{\overline{(\Delta k_{i,t} - \overline{\Delta k_{i,t}})(\Delta k_{i,t-1} - \overline{\Delta k_{i,t-1}})}}{\sigma_t \sigma_{t-1}}. \quad (4)$$

Here, σ_t, σ_{t-1} are the standard deviations of the $\Delta k_{i,t}$ and $\Delta k_{i,t-1}$ distributions, respectively ($\sigma_t \approx \sigma_{t-1}$), and the averaging is performed over all papers in the set. This was done for all k and t. Figure 4 shows that $c_{t,t-1}$ grows with k. For moderately cited papers, $k \ll 60$, the autocorrelation is weak while for highly cited papers, $k \gg 60$, the autocorrelation is strong: $c \sim 1$. The empirical function

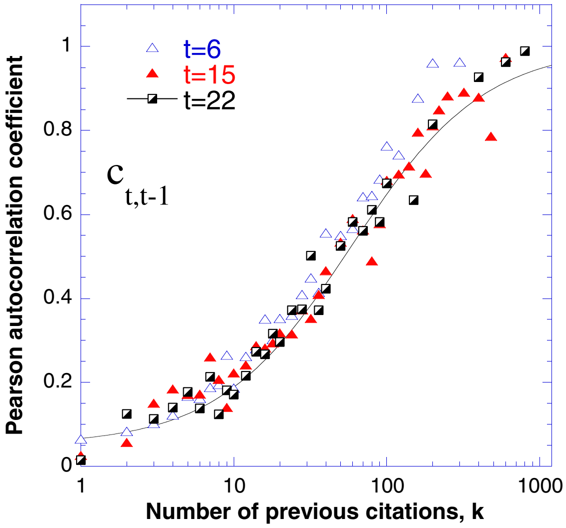

FIG. 4 (color online). The Pearson autocorrelation coefficient for additional citations [Eq. (4)]. Each point corresponds to the set of papers with the same number of previous citations k garnered by a certain citing year t. The data for different t almost collapse. The continuous line shows empirical approximation [Eq. (5)].

$$c(k, t) \approx \frac{k + 3}{k + 60} \quad (5)$$

fits our measurements well. Strong temporal autocorrelation of citations violates the underlying assumption of the preferential attachment model [1,4,5]: it turns out that citations dynamics is not a Markov process since it depends on past history.

We suggest a more realistic growth model that is based on the first-order linear autoregression,

$$\lambda_i = (1 - c)A(k_i + k_0)^\alpha + c\Delta k_{i,t-1}, \quad (6)$$

where λ is the latent citation rate and c is given by Eq. (5). The actual number of additional citations is given by Eq. (3). Equation (6) introduces positive feedback between successive citations of the same paper, in other words, it approximates the citation dynamics of a paper by the inhomogeneous self-exciting point process [17]. (Similar ideas were discussed in Refs. [10,18].) The resulting preferential attachment model replaces Eq. (1) by Eq. (6) in such a way that the stochastic term in Eq. (2) reduces to the Poissonian noise. Equation (6) states that the latent citation rate of a paper [19] depends not only on the total number of accumulated citations but on the recent citation rate as well. This accounts for the "sleeping beauties": the papers that initially had a small number of citations but suddenly became popular. While the conventional preferential attachment model [Eq. (1)] yields predominance of the "first movers" [20], our more realistic model allocates a fair share of citations to "sleeping beauties."

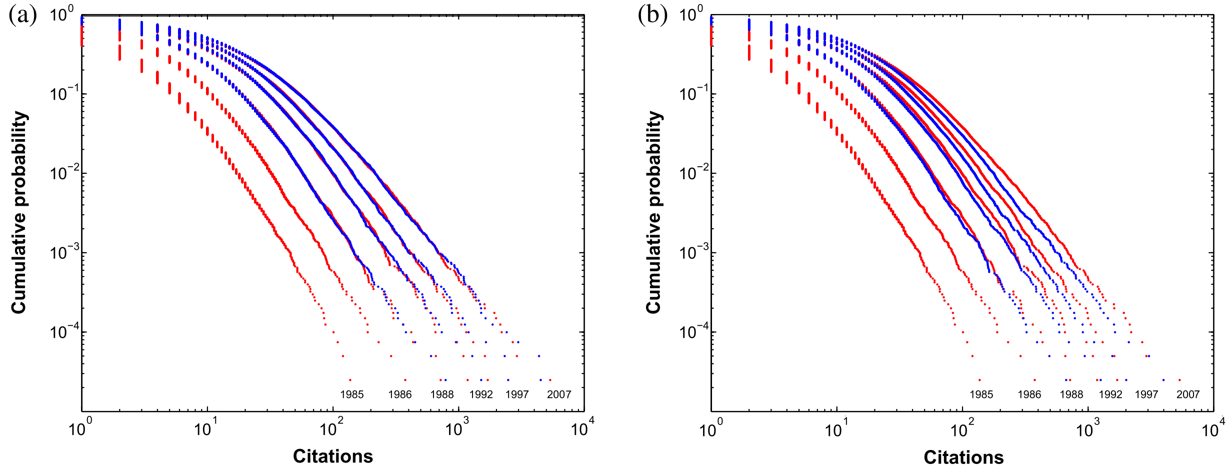

FIG. 5 (color online). Cumulative citation distributions for 40,195 physics papers published in 1984. The citing year is indicated at each curve. Red symbols—measurements, blue symbols—numerical simulation assuming the initial citation distributions of 1985 and 1986. (a) The full model based on Eqs. (3) and (6) provides excellent fits to the measured citation distributions. (b) The incomplete model based on Eq. (1) (correlations ignored) and Eq. (3) underestimates citation counts.

To verify the multiplicative stochastic model described by Eq. (6) we chose all physics papers published in the same year (1984), fixed a certain citing year ($t = 1986$), measured the number of total and last year citations, $k_{i,t}$ and $\Delta k_{i,t-1}$, and calculated λ_i for each paper using Eq. (6) with experimentally measured parameters $c(k)$, $A(t)$, k_0, and $\alpha(t)$. Then we ran numerical simulations assuming Poisson process with the rate given by Eq. (6), found the number of citations of each paper in the year $t + 1$, and calculated the cumulative distribution of citations. The procedure was repeated for the next year and so on. Figure 5(a) shows that this algorithm closely reproduces the actual citation distribution for each citing year. This means that Eq. (6) yields an excellent description not only of the microscopic citation dynamics but of the macroscopic citation distribution as well. On the other hand, the numerical simulation that assumes only the Poisson process and ignores correlations, does not reproduce our measurements well [Fig. 5(b)].

What are the implications of our study? We find that the cumulative citation distribution is neither stable nor stationary but develops in time. Immediately after publication the spread of initial conditions (journal circulation numbers) yields a convex cumulative distribution of citations that can be fitted equally well by the (discrete) power-law [21–23] or log-normal [9,24–26] functions. Thereafter, the citation dynamics of most papers is dominated by the first term in Eq. (6) in such a way that the citation history of papers that managed to garner less than 50–70 citations is completed after 10–15 years. However, the papers with more than 50–70 citations continue to be cited even after 10–15 years, their dynamics being determined by the second term in Eq. (6) which does not decay with time. In other words, while the bulk of the citation distribution becomes stable, the tail grows. In the course of time its shape changes from the convex to concave in such a way that for the most part of the time the tail looks straight in the log-log coordinates. Although such a power-law tail was previously considered as a fingerprint of the scale-free network, at least for the citation network it turns out to be a transient phenomenon. The intrinsic scale of the citation network, $k_{\mathrm{cr}} = 50$–70, is clearly revealed in the microscopic dynamics (Fig. 4). We conclude that the almost power-law degree distribution of citations that was previously interpreted as the indication of the scale-free network [1,3,5,21] arises from the interplay between aging [27], multiplicative stochastic process [Eq. (2)], and superlinear preferential attachment.

The two-term Eq. (6) implies that scientific papers constitute two broad classes with respect to their longevity [9]. The citation rate of 90% of the papers achieves its maximum in 2–3 years after publication and decays to zero in 10–15 years. The citation dynamics of these papers is the aftereffect of their initial hit and is more or less predictable since the impact of these papers is probably limited to several research groups and does not propagate further. However, the citation rate of 10% of the papers that overcome the tipping point [23] of $k_{\mathrm{cr}} \simeq 50$–$70$ citations is determined more by their recent citation history. It seems that these papers have a continuing impact [28] which propagates from one research group to another in a cascade process like in epidemics [29]. This diffusion of scientific knowledge [30] extends the paper longevity to much more than 10–15 years.

In summary, our measurements indicate that the mechanism that generates complex networks may be more sophisticated than the memoryless linear preferential attachment assumed so far. We propose a stochastic growth model that considers the evolution of the node degree as an inhomogeneous self-exciting point process. In the context of citations, the model is fully verified by our microscopic and macroscopic measurements and can serve for prognostication of the future citation behavior of a paper, group of papers, or of a journal's impact factor.

We are grateful to S. Redner, N. Shnerb, and A. Scharnhorst for insightful discussions.

*michael.golosovsky@mail.huji.ac.il

[1] R. Albert and A.L. Barabasi, Rev. Mod. Phys. **74**, 47 (2002).

[2] S.N. Dorogovtsev and J.F.F. Mendes, Adv. Phys. **51**, 1079 (2002).

[3] M.E.J. Newman, SIAM Rev. **45**, 167 (2003); Phys. Today **61**, No. 11, 33 (2008).

[4] H.A. Simon, Biometrika **42**, 425 (1955).

[5] D. de Solla Price, Science **149**, 510 (1965); J. Am. Soc. Inf. Sci. **27**, 292 (1976).

[6] P.L. Krapivsky, S. Redner, and F. Leyvraz, Phys. Rev. Lett. **85**, 4629 (2000).

[7] H. Jeong, Z. Neda, and A.L. Barabasi, Europhys. Lett. **61**, 567 (2003).

[8] Y.-H. Eom, C. Jeon, H. Jeong, and B. Kahng, Phys. Rev. E **77**, 056105 (2008).

[9] S. Redner, Phys. Today **58**, No. 6, 49 (2005).

[10] M. Wang, G. Yu, and D. Yu, Physica (Amsterdam) **387A**, 4692 (2008).

[11] A. Scharnhorst, K. Borner, and P. van den Besselaar, *Models of Science Dynamics* (Springer, Berlin, 2012).

[12] G. Bianconi and A.L. Barabasi, Phys. Rev. Lett. **86**, 5632 (2001).

[13] M. Medo, G. Cimini, and S. Gualdi, Phys. Rev. Lett. **107**, 238701 (2011).

[14] S. Valverde, R.V. Sole, M.A. Bedau, and N. Packard, Phys. Rev. E **76**, 056118 (2007).

[15] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.109.098701 for the methodology.

[16] G. Csardi, K.J. Strandburg, L. Zalanyi, J. Tobochnik, and P. Erdi, Physica (Amsterdam) **374A**, 783 (2007).

[17] A.G. Hawkes, Biometrika **58**, 83 (1971).

[18] C. Cattuto, V. Loreto, and V.D.P. Servedio, Europhys. Lett. **76**, 208 (2006).

[19] Q.L. Burrell, J. Am. Soc. Inf. Sci. **54**, 372 (2003).

[20] M.E.J. Newman, Europhys. Lett. **86**, 68001 (2009).

[21] S. Redner, Eur. Phys. J. B **4**, 131 (1998).

[22] M.L. Wallace, V. Lariviere, and Y. Gingras, J. Informetrics **3**, 296 (2009).

[23] G.J. Peterson, S. Presse, and K.A. Dill, Proc. Natl. Acad. Sci. U.S.A. **107**, 16023 (2010).

[24] A.M. Petersen, F. Wang, and H.E. Stanley, Phys. Rev. E **81**, 036114 (2010).

[25] M.J. Stringer, M. Sales-Pardo, and L.A.N. Amaral, PLoS ONE **3**, e1683, (2008).

[26] F. Radicchi, S. Fortunato, and C. Castellano, Proc. Natl. Acad. Sci. U.S.A. **105**, 17268 (2008).

[27] S.N. Dorogovtsev and J.F.F. Mendes, Phys. Rev. E **62**, 1842 (2000).

[28] M. Golosovsky and S. Solomon, Eur. Phys. J. **205**, 303 (2012).

[29] W. Goffman and V.A. Newill, Nature (London) **204**, 225 (1964).

[30] A.B. Jaffe and M. Trajtenberg, Proc. Natl. Acad. Sci. U.S.A. **93**, 12671 (1996).