PHYSICAL REVIEW E **95**, 012324 (2017)

Growing complex network of citations of scientific papers: Modeling and measurements

Michael Golosovsky* and Sorin Solomon

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

(Received 4 May 2016; revised manuscript received 1 November 2016; published 30 January 2017)

We consider the network of citations of scientific papers and use a combination of the theoretical and experimental tools to uncover microscopic details of this network growth. Namely, we develop a stochastic model of citation dynamics based on the copying-redirection-triadic closure mechanism. In a complementary and coherent way, the model accounts both for statistics of references of scientific papers and for their citation dynamics. Originating in empirical measurements, the model is cast in such a way that it can be verified quantitatively in every aspect. Such validation is performed by measuring citation dynamics of physics papers. The measurements revealed nonlinear citation dynamics, the nonlinearity being intricately related to network topology. The nonlinearity has far-reaching consequences including nonstationary citation distributions, diverging citation trajectories of similar papers, runaways or "immortal papers" with infinite citation lifetime, etc. Thus nonlinearity in complex network growth is our most important finding. In a more specific context, our results can be a basis for quantitative probabilistic prediction of citation dynamics of individual papers and of the journal impact factor.

DOI: 10.1103/PhysRevE.95.012324

I. INTRODUCTION

Complex networks became objects of physics research after appearance of the Internet, large information databases, and mapping of genetic and metabolic networks. The research in networks focused initially on network topology [1,2] while presently it focuses more on temporal and evolving networks [3] and dynamic processes such as network growth. The paradigm for complex network growth is the cumulative advantage mechanism invented by de Solla Price [4]. The most quantified complex network in his time was citation network which exhibited an intriguing power-law degree distribution which was considered as evidence of the scale-free behavior. de Solla Price sought to explain it, so he postulated that citation network grows by addition of new papers that cite older papers with probability

$$\lambda_i \propto (K_i + K_0), \tag{1}$$

where K_i is the number of citations of the target paper i and K_0 is an unspecified constant. de Solla Price showed that the linear growth rule captured by Eq. (1) generates networks with the power-law degree distribution. With appearance of the Internet and vigorous advent of network science, a similar rule was invented by Barabasi [5] who suggested that Eq. (1) is the most generic growth rule of complex networks. The Barabasi-Albert model or preferential attachment is also known colloquially as a "rich get richer" or Matthew effect [6]. Equation (1) was soon generalized to include aging and nonlinearity [7,8],

$$\lambda_i = A(t)[K_i + K_0]^{\delta}. \tag{2}$$

Here, $A(t)$ is the aging function, common to all nodes, K_0 is the initial attractivity, and δ is the growth exponent. The measurements on many complex networks [6] verified Eq. (2) and showed ubiquity of networks with $\delta \sim 1$.

Although Eq. (2) successfully describes the complex network growth, it is associated with several conceptual difficulties. Indeed, this equation encodes an empirical rule assuming that each node in the network garners new links with the rate proportional to its current degree; in other words, Eqs. (1) and (2) assume that all nodes differ only in one dimension—degree. This assumption results in similar growth dynamics of the nodes of the same age, while in reality there is a huge diversity in their growth trajectories.

To solve this difficulty, Bianconi and Barabasi [9] added a different dimension to node description—fitness. This notion replaced the egalitarian picture, according to which all nodes are born equal, by the picture where each node is born with some intrinsic propensity of growth. The corresponding growth rule [10] (see also Refs. [11,12]) becomes

$$\lambda_i = \eta_i A(t)[K_i + K_0], \tag{3}$$

where η_i is the node fitness—an empirical parameter introduced on top of the preferential attachment. To be less empiric, several authors [13–16] added a more physical sense to Eq. (3) and replaced η_i by node similarity (homophily). The latter notion captures the fact that a new node tends to link to the nodes with similar content rather than to randomly chosen nodes. Technically, this line of reasoning results in Eq. (3) where η_i is replaced by η_{ij}, the latter quantifying the similarity between the two connecting nodes [17].

Still, Eq. (3) contains too many empirical parameters that prompt for microscopic explanation. The need for such explanation becomes even more evident after realizing that Eqs. (1)–(3) are *global*. In order for a new node to attach preferentially to most popular nodes it shall be familiar with the whole network. This global picture is unrealistic and many efforts have been spent to elucidate the *local* microscopic mechanism staying behind Eqs. (1)–(3).

The most popular local mechanism is the copying rule [18] which is also known as recursive search [19], link copying or redirection [8,20,21], random walk or local search [22,23], triple or triangle formation [24], transitive triples [25], or triadic closure [26]. A similar rule operates in social networks [22,27], propagation of ideas [28–30], diffusion of innovations

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

[31], and citation dynamics [32]. This rule assumes that a new node performs random and recursive searches: first, it attaches to a randomly chosen node, second, it copies some links of the latter. This results in the following dynamic equation:

$$\lambda_i = A(t)[cK_i + (1-c)K_0], \quad (4)$$

where $A(t)$ is the aging factor, the first and the second addends in the parentheses correspond to the recursive and random searches, respectively, and the parameter c regulates the relative weights of the two. Similar two-term growth equations were suggested by Refs. [13,27,33,34]. Equation (4) is formally identical to Eq. (1) in which K_0 captures the probability of random search. The intuition behind the first addend in Eq. (4) is as follows: if some node i has K_i links, the probability to find it through recursive search is increased by a factor K_i. Thus Eq. (4) seems to provide a natural explanation for the preferential attachment mechanism. Indeed, Refs. [35,36] rigorously demonstrated that Eq. (1) can evolve from the copying rule.

However, the parameters of Eq. (4) were never measured systematically: it is not known whether time dependencies of the random and recursive search are the same or differ, whether the probability of recursive search is identical for all nodes of the same age or not. Our goal is to measure dynamic parameters of some complex network, to establish its microscopic growth rules, and to compare them to existing models. To this end we consider an iconic example of a growing network—citations to scientific papers—having in mind that the models of network growth were originally suggested in relation to this very network [37]. Despite some specificity (it is ordered, acyclic, and does not allow rewiring and link deletion), citation network is a well-documented prototypical directed network. Following Ref. [38] we adopt a comprehensive approach, namely, we consider the network growth from two perspectives: that of an author and that of a cited paper. The former approach focuses on the composition of the reference list of a paper, the latter one focuses on the papers that cite a given paper. We establish duality between these two approaches and formulate a stochastic model that accounts both for citation dynamics of scientific papers and for the age composition of their reference lists.

The paper is organized as follows. Section II focuses on references. We propose there a plausible scenario that the authors follow when they compose the reference lists of their papers. On the basis of this scenario we develop the model accounting for the age distribution of references in the reference list of scientific papers. The model contains empirical functions which we find in dedicated measurements. Section III focuses on the reference-citation duality. We develop here a mean-field model of citation dynamics which is based on this duality and on the model for age distribution of references introduced in Sec. II. This macroscopic approach captures the mean citation dynamics of a single research field. In Sec. IV we develop an "individualized" mean-field model that captures citation dynamics of the groups of similar papers. This mesoscopic approach focuses on the deterministic component of citation dynamics and leaves out its stochastic component. Section V deals with stochastic model of citation dynamics of individual papers and its validation. This represents a truly microscopic approach and it is the main message of the paper. The readers who want to skip intermediate steps can read Sec. II A (scenario), Sec. IV A (definitions), and jump to Sec. V. The logical flow of the paper is discussed in more detail in the Supplemental Material (SM I) [39].

II. CITATION DYNAMICS FROM THE AUTHORS' PERSPECTIVE

We discuss here how the authors compose the reference lists of their papers, then we formulate a model of the age composition of the reference list of papers and calibrate it in measurements with physics papers.

A. Recursive search algorithm

The composition of the reference lists of scientific papers is the clue to citation analysis. While citation dynamics of a paper is determined by several factors: popularity of the research field, journal impact factor, preferences and tastes of citing authors, etc., the reference list derives from only one source: decision of an author (research team) who chooses the references on the basis of their content and age. We focus here on the age of the references and do not consider their content, although this can be very important [13].

Our goal is to measure and to model the age composition of the reference lists of papers. To this end we distinguish between two kinds of references: the direct references are those that are not cited by any other paper in the reference list of the source paper, the indirect references are those cited by one or several preselected references (see Fig. 1).

What is the source of indirect references? If the author cites some old seminal studies, his most recent references will probably cite them as well. In our parlance these old papers are indirect references. On the other hand, indirect references may result from copying. Indeed, consider an author who writes a

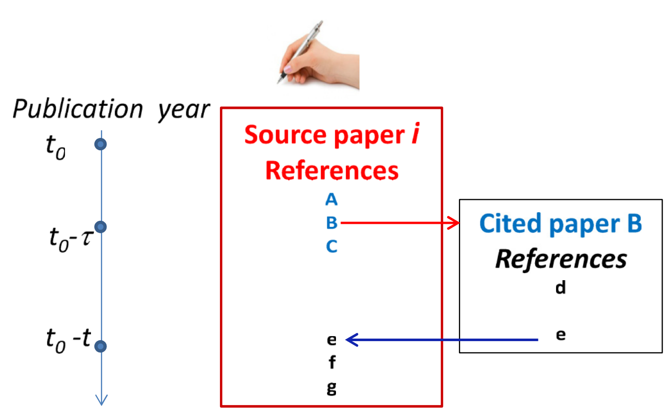

FIG. 1. Cartoon scenario of the referencing process which is the basis of our model. Consider a paper i published in the year t_0 and its list of references $A, B, C \dots e, f, g$ arranged in descending chronological order. The author of i found some references (such as A, B, C) independently and copied some others (such as e) from the reference lists of already selected papers (for example, from B). We assume that the probability of the paper e to be copied into reference list of i is determined by the time lag τ between publication years of its parent paper B and of its grandparent paper i.

research paper. He reads scientific journals or media articles, searches databases, finds relevant papers, and includes some of them in his reference list. These are direct references [40]. Then he studies the reference lists of these papers, picks up relevant references, reads them, and adds some of them into his reference list. These are indirect references. Then he studies the reference lists of the newly added papers, copies some references, and continues recursively. All those papers that were found through reference lists of the already selected papers are indirect references. In what follows we analyze the age composition of the reference list of papers generated by this copying (recursive search) mechanism. Our analysis is based on the causality principle that requires the indirect references to be older than their preselected sources.

B. Model: Age distribution of references

To quantitatively account for the age composition of the reference lists of research papers we develop an analytical model based on the causality principle and recursive search algorithm. Consider a source paper i published in the year t_0 and one of its references B published in the year $t_0 - \tau$ (Figs. 1 and 2). We assume that once i cites B, it can copy any paper e from its reference list with equal probability. This probability depends on a variety of factors, such as the local structure of citation network, the age difference between the papers i, B, and e, and their similarity (although the latter is already captured by the parental relations: e is a descendant of i).

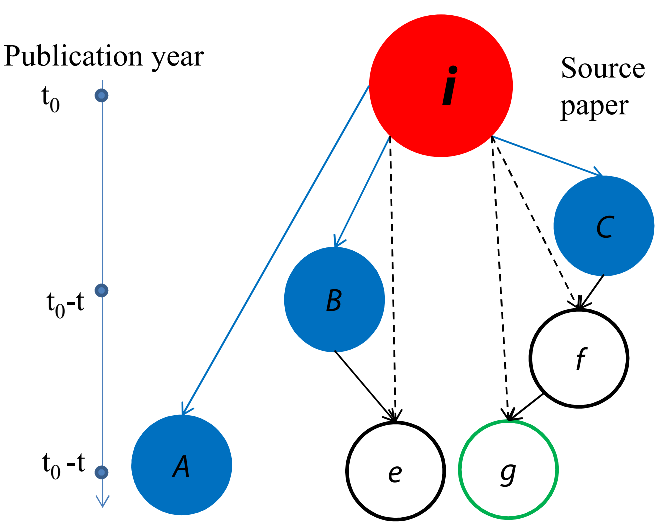

FIG. 2. Direct and indirect references. Consider a source paper i published in the year t_0 and its list of references $A,B,C \ldots e,f,g$ arranged in descending chronological order. The papers A,B,C are not cited by any other papers in the reference list of i and they are direct references. The papers e and f are copied from the references lists of the direct references B and C, correspondingly, and they are indirect references. The paper g is copied from the reference list of f and it is also an indirect reference. Each indirect reference closes a triangle in which the source paper i is one of the vertices. The solid and dashed lines connect, correspondingly, the direct and indirect references to their parent papers. The references published in each year include direct and indirect references. In particular, the references published in the year $t_0 - t$ consist of A, e, and g.

Following Ref. [24] we develop this scenario into an analytical model accounting for the average age composition of the reference lists of papers. Indeed, consider a set of papers in one scientific field that were published in one year t_0. We denote by $R(t_0,t_0 - t)$ the average number of references in the reference list of these papers that were published in the year $t_0 - t$. These consist of the direct and indirect references,

$$R(t_0,t_0 - t) = R_{\mathrm{dir}}(t_0,t_0 - t) + R_{\mathrm{indir}}(t_0,t_0 - t). \quad (5)$$

The function $R_{\mathrm{dir}}(t_0,t_0 - t)$ is exogenous to our model. Once it is known, the model calculates $R_{\mathrm{indir}}(t_0,t_0 - t)$. To find $R_{\mathrm{indir}}(t_0,t_0 - t)$ we consider a source paper i published in the year t_0 and one of its references B published in the year $t_0 - \tau$ (Fig. 2). The reference list of the latter contains $R(t_0 - \tau,t_0 - t)$ references published in the year $t_0 - t$. We assume that each of them can be copied to the reference list of the source paper i with equal probability $P(t_0,t_0 - \tau)$ which is an empirical time-dependent function. $P(t_0,t_0 - \tau)$ is the probability that the second-generation reference picked up from the first-generation reference that was published in the year $t_0 - \tau$ is copied to the reference list of the source paper which was published in the year t_0. The number of indirect references in the reference list of i that were published in the year $t_0 - t$ is the sum of contributions made by all references of i that were published earlier,

$$R_{\mathrm{indir}}(t_0,t_0 - t) = \sum_{\tau=0}^{t} R(t_0 - \tau,t_0 - t)P(t_0,t_0 - \tau)R(t_0,t_0 - \tau). \quad (6)$$

Figure 2 visualizes these reference cascades. At the head of a cascade there is a direct reference. It entails an indirect reference that can entail another indirect reference and so on. All these cascades could have been captured by Eq. (6) if instead of $R(t_0,t_0 - \tau)$ we were using $R_{\mathrm{dir}}(t_0,t_0 - \tau)$ in the kernel. Nevertheless, we prefer to write Eq. (6) with $R(t_0,t_0 - \tau)$ as a source, since in this form Eq. (6) embodies the recursive search algorithm more straightforwardly.

So far, Eq. (6) does not contain new information since almost any function $R(t)$ can be decomposed into the sum of two functions $R_{\mathrm{dir}}(t)$ and $R_{\mathrm{indir}}(t)$ that satisfy Eqs. (5) and (6) (if one chooses the appropriate kernel P). While we do not make any statement with respect to the function $R_{\mathrm{dir}}(t_0,t_0 - t)$, our model assumes that the kernel $P(t_0,t_0 - \tau)$ has a simple functional form reflecting the author's psychology. In what follows we measure the functions $R_{\mathrm{dir}}(t_0,t_0 - \tau), R_{\mathrm{indir}}(t_0,t_0 - \tau)$, and $R(t_0,t_0 - \tau)$, solve integral Eq. (6), and determine $P(t_0,t_0 - \tau)$. We find that the latter function is indeed a simple exponential which can be conveniently interpreted in the framework of the copying mechanism. Thus our measurements of references validate the copying mechanism and justify it as a foundation of the model of citation dynamics.

C. Measurements: Direct and indirect references

1. Reduced age distribution of references

To develop our measurement protocol we note that the age distribution of references $R(t_0,t_0 - t)$ depends on two variables: t_0 and t. To exclude dependence on the publication year t_0 we follow Refs. [41–48] and consider the reduced age

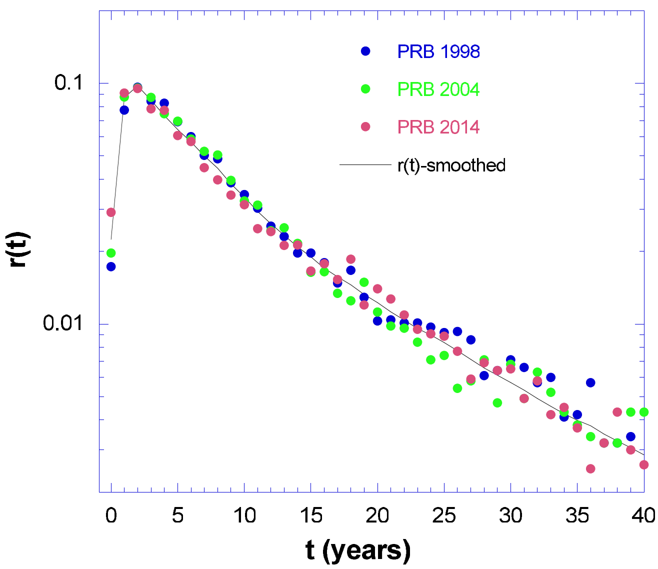

FIG. 3. $r(t)$, reduced age distribution of references (a fraction of references published in the year $t_0 - t$ that appear in the reference list of a paper published in the year t_0). Red, blue, and green circles stay for three sets of research papers published in July issues of *Physical Review B* in 1998, 2004, and 2014, correspondingly. Similar to previous studies [41,44], we observe that $r(t)$ dependencies for all publication years collapse onto a single curve (with possible exception of $t = 0$). Continuous line was obtained by averaging and smoothing the data.

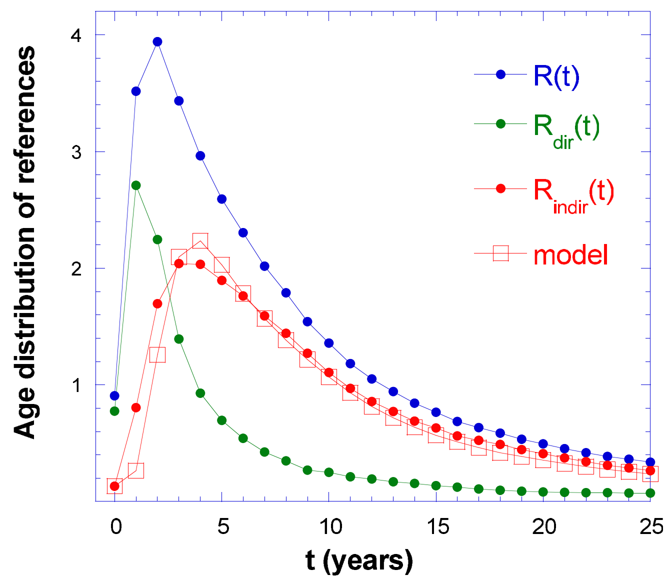

FIG. 4. Time dependence of $R_{\rm dir}$, $R_{\rm indir}$, and $R = R_{\rm dir} + R_{\rm indir}$, the numbers of direct, indirect, and total references in a typical reference list of the *Physical Review B* papers. The data show the average values for 21 PRB papers published in 2014. Empty squares show model prediction based on Eq. (6) with exponential kernel $P(\tau) = P_0 e^{-(\gamma+\beta)\tau}$ where $P_0 = 0.19$ and $\gamma + \beta = 1.2\ {\rm yr}^{-1}$.

distribution of references

$$r(t) = \frac{R(t_0,t_0 - t)}{R_0(t_0)}, \tag{7}$$

where $R_0(t_0)$ is the average reference list length of the papers published in the year t_0. Figure 3 shows $r(t)$ dependence. After a sharp growth for $t = 0$–2 it slowly decays with t as $R(t) \sim \frac{1}{(t+0.2)^{1.5}}$ (the publication year corresponds to $t = 0$).

We note a remarkable fact: $r(t)$ is almost independent of the publication year t_0. Therefore, we can write $R(t_0,t_0 - t) = r(t)R_0(t_0)$. This means that $R(t_0,t_0 - t)$ dependence on t_0 results only from the $R_0(t_0)$ dependence. Since R_0 grows with time exponentially, $R_0(t_0) \propto e^{\beta t_0}$ (SM III [39]), then Eq. (6) can be recast as follows:

$$R_{\rm indir}(t_0,t_0 - t) = \sum_{\tau=0}^{t} R(t_0,t_0 - t + \tau)P(t_0,t_0 - \tau) \times e^{-\beta\tau} R(t_0,t_0 - \tau). \tag{8}$$

Now all functions in Eq. (8) belong to the same publication year t_0. Hence, they can be considered as functions of only one independent variable—t. We cut short our notation and instead of $R(t_0,t_0 - t), R_{\rm indir}(t_0,t_0 - t)$ and $P(t_0,t_0 - t)$ we write $R(t)$, $R_{\rm indir}(t)$, and $P(t)$, correspondingly. Thus we come to a compact expression

$$R_{\rm indir}(t) = \sum_{\tau=0}^{t} R(t - \tau)P(\tau)e^{-\beta\tau} R(\tau). \tag{9}$$

Although the kernel of Eq. (9) includes the product of $R(t - \tau)$ and $R(\tau)$ this is still a linear equation [as well as Eqs. (6) and (8)]. The reason is that $R_{\rm indir}(t) \propto R_0$ and $R(t) \propto R_0$, but $P(t) \propto R_0^{-1}$ (we will show this in Sec. II C 3), in such a way that in fact both sides of Eq. (8) linearly depend on the number of references.

2. Measurement of $R_{\rm dir}(t), R_{\rm indir}(t)$

In what follows we measure $R(t)$, $R_{\rm dir}(t)$, and $R_{\rm indir}(t)$, compare our results to Eq. (8), and determine $P(\tau)$. Figure 4 shows the results of such measurements for a set of 21 physics papers published in *Physical Review B* in 2014 (SM II [39]). $R_{\rm dir}(t)$ sharply increases during first couple of years after publication and then slowly decays while $R_{\rm indir}(t)$ at first slowly increases and then decays even more slowly. Do these observations make sense from the viewpoint of the copying model (Fig. 1)? Reference [49], which was the first to suggest the decomposition of the reference list of a paper into direct and indirect contributions, assumed that $R_{\rm dir}(t)$ decays fast. This extreme picture reduces the referencing process to ridicule: the author reads only few recent papers and copies all other references from them. Our measurements reveal a much more realistic scenario of the referencing process. Indeed, the long tail of $R_{\rm dir}(t)$ implies that the author reads the recent and old papers as well, and copies only few references from each of them.

To find the kernel $P(\tau)$ we assumed exponential time dependence $P(\tau) = P_0 e^{-\gamma\tau}$. We substituted into Eq. (8) this kernel and the measured function $R(t)$. By varying fitting parameters P_0 and γ we searched for the best correspondence between the calculated and the measured $R_{\rm indir}(t)$ dependencies. Figure 4 shows that it occurs for $P_0 = 0.19$ and $\gamma + \beta = 1.2\ {\rm yr}^{-1}$ (SM VI [39]). Fast exponential decrease of $P(\tau)$ suggests that if the references are copied, this is done preferably from recent references, as expected. The large proportion of indirect references in the reference list of papers as implied by a rather big P_0 (we found that the average reference list of a *Physical Review B* paper includes 65% indirect references) conforms well with previous estimates of Refs. [15,23,49–51] (see also SM V [39]). Thus our findings support the copying mechanism.

3. Dependence on the publication year

In what follows we explore how the functions $R_{\rm dir}(t_0,t_0-t)$, $R_{\rm indir}(t_0,t_0-t)$, and $P(t_0,t_0-t)$ depend on the publication year t_0. To this end we introduce $r_{\rm dir}(t)=\frac{R_{\rm dir}(t_0,t_0-t)}{R_0(t_0)}$ and $r_{\rm indir}(t)=\frac{R_{\rm indir}(t_0,t_0-t)}{R_0(t_0)}$ and recast Eq. (5) as follows:

$$r(t)=r_{\rm dir}(t)+r_{\rm indir}(t). \tag{10}$$

Since $r(t)$ does not depend on the publication year t_0 (Fig. 3), we make a plausible assumption that $r_{\rm dir}(t)$ and $r_{\rm indir}(t)$ do not either, in such a way that $R_{\rm dir}(t_0,t_0-t)$ and $R_{\rm indir}(t_0,t_0-t)$ dependencies on t_0 reduce to $R_0(t_0)$ dependence. This allows making certain conclusions on how $P(t_0,t_0-t)$ depends on the publication year t_0. To this end we divide both parts of Eq. (8) by $R_0(t_0)$ and come to

$$r_{\rm indir}(t)=\sum_{\tau=0}^{t} r(t-\tau)T(t_0,t_0-\tau)r(\tau), \tag{11}$$

where $T(t_0,t_0-\tau)=P(t_0,t_0-\tau)R_0(t_0-\tau)=T_0e^{-(\gamma+\beta)\tau}$ and $T_0=P_0(t_0)R_0(t_0)$.

Since neither $r_{\rm indir}(t)$ nor $r(t)$ depend on t_0, then T_0 should not depend on t_0 either. We infer from Fig. 4 that $T_0=7.6$, $\gamma+\beta=1.2\ {\rm yr}^{-1}$. This means that the author of a paper published in the year t_0 copies on average 7.6 references from each preselected reference published in the same year, 2.3 references from each one-year-old preselected reference, 0.7 references from each two-year-old preselected reference, and so on.

With respect to the probability of indirect citation P, we note that $P_0(t_0)=\frac{T_0}{R_0(t_0)}$. While T_0 does not depend on the publication year, R_0 slowly increases with growing t_0 (SM III [39]). This means that P_0 decreases with t_0.

III. REFERENCE-CITATION DUALITY

Our further task is the extension of the recursive search (copying) model (Fig. 1) to citation dynamics. This can be done by two complementary approaches: reformulation of the model from the perspective of a citing paper or by exploration of the reference-citation duality. We focus here on the latter approach and leave the former one for the next section, Sec. IV. In this section we develop the mean-field model of citation dynamics for the papers in the same field published in one year.

A. Duality

Scientific papers represent a directed complex network. Figure 5 shows a part of this network consisting of two sets of papers published in the years t_0 and t_0+t, correspondingly. The papers are nodes while the links between them can be considered either as citations or references since one paper's citation is another paper's reference. To explore mathematical consequences of this duality we introduce $N_{\rm publ}(t_0)$ and

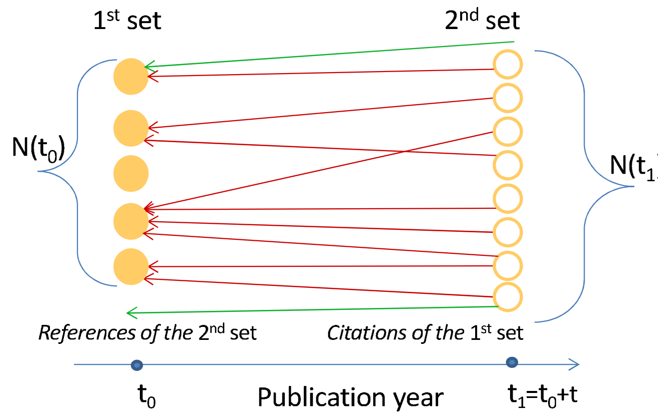

FIG. 5. Reference-citation duality. The filled and the empty circles show all papers in one research field that were published in the years t_0 and $t_1=t_0+t$, correspondingly. The links between the two sets are shown by red lines. With respect to the first set these links are citations, with respect to the second set they are references. Green lines show interdisciplinary citations and references.

$N_{\rm publ}(t_0+t)$, the number of papers in each set; $M(t_0,t_0+t)$, the mean number of citations garnered in the year t_0+t by a paper of the first set published in t_0; and $R(t_0+t,t_0)$, the average number of references published in the year t_0 that appear in the reference list of the papers of the second set which is published in the year $t_1=t_0+t$. We assume that all citing papers belong to the same research field and neglect interdisciplinary papers, books, and other references or citations which are not research papers. Under this assumption, the number of papers that cite the first set and that were published in the year t_0+t is equal to the number of references published in the year t_0 that appear in the reference lists of the papers of the second set, namely,

$$N_{\rm publ}(t_0)M(t_0,t_1)=N_{\rm publ}(t_1)R(t_1,t_0). \tag{12}$$

Since the annual growth of the number of publications is nearly exponential, $N_{\rm publ}(t_0+t)\approx N_{\rm publ}(t_0)e^{\alpha t}$ (SM III [39]), Eq. (12) yields

$$M(t_0,t_0+t)=R(t_0+t,t_0)e^{\alpha t}. \tag{13}$$

We replace here $R(t_0+t,t_0)$ by $r(t)R_0(t_0+t)=r(t)R_0(t_0)e^{\beta t}$ where $r(t)$ is the reduced distribution of references (Sec. II C 1). Since $r(t)$ does not depend on the publication year, then $R(t_0+t,t_0)=R(t_0,t_0-t)\frac{R_0(t_0+t)}{R_0(t_0)}=R(t_0,t_0-t)e^{\beta t}$. We substitute this expression into Eq. (13) and find

$$M(t_0,t_0+t)=R(t_0,t_0-t)e^{(\alpha+\beta)t}. \tag{14}$$

Equation (14) captures the reference-citation duality. It relates synchronous (retrospective) and diachronous (prospective) citation distributions [41–48] for the same publication year. To compare these distributions for different publication years t_0 and t_1 one only needs to introduce a constant factor $\frac{R_0(t_1)}{R_0(t_0)}$, in such a way that $R(t_1,t_1-t)=R(t_0,t_0-t)\frac{R_0(t_1)}{R_0(t_0)}$, $M(t_1,t_1+t)=M(t_0,t_0+t)\frac{R_0(t_1)}{R_0(t_0)}$.

Figure 6 validates Eq. (14) and proves that $M(t_0,t_0+t)$ and $R(t_1,t_1-t)$ are related to one another through the exponential factor $e^{(\alpha+\beta)t}$. Although this factor grows very slowly with time, it is responsible for a subtle qualitative difference

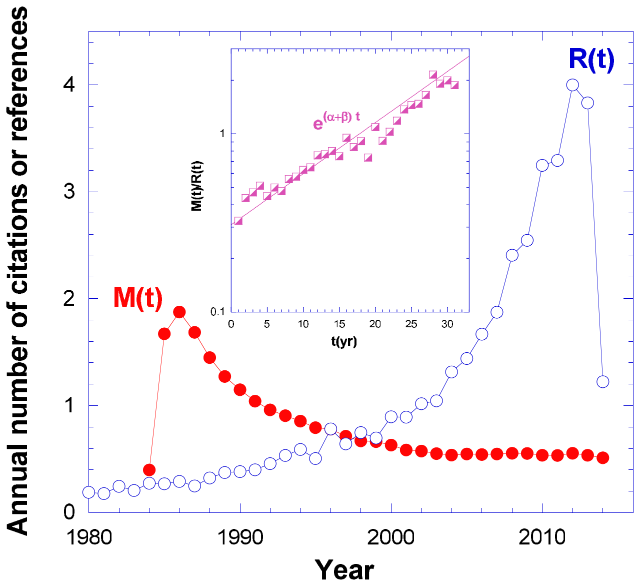

FIG. 6. Reference-citation duality. $M(t_0,t_0+t)$ is mean annual number of citations of 48 168 physics papers published in $t_0 = 1984$. $R(t_1,t_1-t)$ is the age composition of the reference list of the *Physical Review B* papers published in $t_1 = 2014$ (the data are from Fig. 3). Both dependencies are qualitatively similar and almost obey mirror symmetry. The inset shows the ratio of $M(t_0,t_0+t)$ to $R(t_1,t_1-t)$. The straight line indicates exponential dependence $e^{(\alpha+\beta)t}$ suggested by Eq. (14) with $\alpha = 0.046\ \mathrm{yr}^{-1}$ and $\beta = 0.02\ \mathrm{yr}^{-1}$ as found in SM III [39].

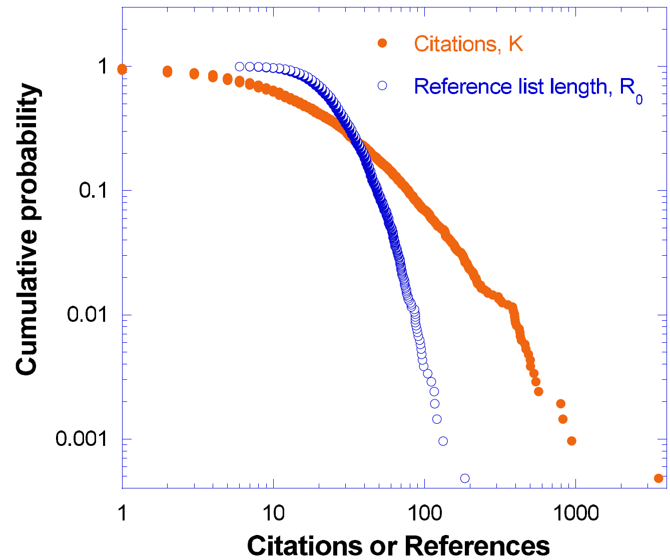

FIG. 7. Cumulative distribution of the reference list lengths R_0 and of the long-time limit of citations K_∞ for the same set of papers (all 2078 *Physical Review B* papers published in 1984). Citations were counted in 2014. While both distributions have almost the same mean, R_0 exhibits a relatively narrow bell-shaped distribution while K_∞ distribution is very wide and has a fat tail.

between $M(t_0,t_0+t)$ and $R(t_1,t_1-t)$ dependencies. Indeed, we infer from Figs. 3 and 6 that the integral $\int_0^t R(t_1,t_1-\tau)d\tau$ converges to $R_0(t_1)$ as $t\to\infty$. However, the function $M(t_0,t_0+t)$ decays slower due to exponential factor $e^{(\alpha+\beta)t}$. Thus, the integral $\int_0^t M(t_0,t_0+\tau)d\tau$ can diverge as $t\to\infty$. This is exactly the situation with physics papers.

B. Mean-field citation dynamics

The reference-citation duality naturally leads to a dynamic equation for $M(t_0,t_0+t)$, an average number of citations garnered by a paper in the year t after publication, where the averaging is performed over all papers in one field published in one year. Indeed, we represent $M(t_0,t_0+t)$ as a sum of direct and indirect citations,

$$M(t_0,t_0+t) = M_{\mathrm{dir}}(t_0,t_0+t) + M_{\mathrm{indir}}(t_0,t_0+t), \quad (15)$$

replace t_0 by t_0+t in Eq. (6), substitute there R by M using Eq. (14), and arrive at

$$M_{\mathrm{dir}}(t_0,t_0+t) = r_{\mathrm{dir}}(t)R_0(t_0)e^{(\alpha+\beta)t}, \quad (16)$$

$$M_{\mathrm{indir}}(t_0,t_0+t) = \sum_{\tau=0}^{t} M(t_0+\tau,t_0+t)P(t_0+t,t_0+\tau) \times M(t_0,t_0+\tau). \quad (17)$$

Although dynamic equations for citations [Eqs. (15) and (17)] are very similar to Eqs. (5) and (6) for references, there is a dramatic difference in their statistics (in and out degrees in network language). Figure 7 shows that citation distribution is extremely broad, while the reference list length distribution for the same set of papers is a relatively narrow bell-shaped curve. (The World Wide Web exhibits a similar asymmetry between in- and out-degree distributions [52].) Narrow R_0 distribution implies that $R(t)$ truly represents the age composition of the reference list of an average paper. Broad citation distribution indicates that Eqs. (15) and (17) describe citation dynamics only in the mean-field approximation; citation dynamics of individual papers can be qualitatively different from the mean.

What is the source of asymmetry between R_0 and K_∞ distributions? To our opinion, this derives from the fact that references are compiled by one author while citations derive from many authors. Indeed, while journal style requirements do not standardize the reference list length, the authors try to comply with what is accepted in their research field. This means that there is a feedback mechanism that forces the authors to adhere to some average reference list length and this results in a relatively narrow distribution of R_0. However, if we consider citation dynamics of a paper, the decision on whether to cite it comes from many uncoordinated authors. There is no feedback mechanism regulating the number of citations of a paper and this is the source of enormous variability in citation dynamics of individual papers.

All these considerations suggest that the modeling of the citation dynamics of individual papers requires a special approach. However, any such approach shall be compatible with Eqs. (15) and (17) which capture citation dynamics of papers in the mean-field approximation. In what follows we develop a model of citation dynamics based on recursive search mechanism (Fig. 1) that satisfies this requirement.

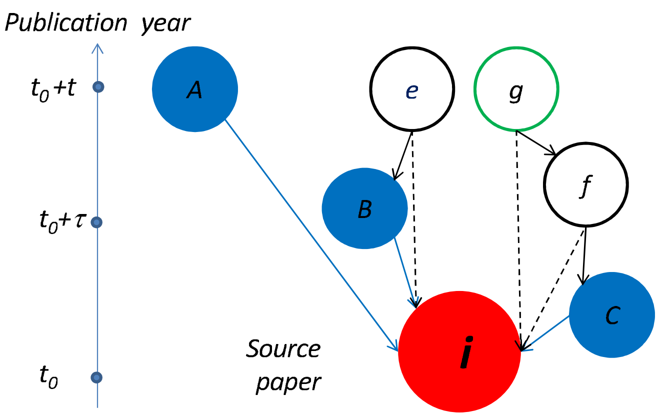

FIG. 8. A fragment of a citation network showing a source paper i and its citing papers. The papers A,B,C are direct citations since they cite i and do not cite any other paper citing i. The papers e,f cite papers B,C, correspondingly, and they are indirect citations. The paper g is also an indirect citation since it cites f that cites i. The solid and dashed lines link the source paper with its direct and indirect citing papers, correspondingly. Each indirect citation closes a triangle in which the source paper i is one of the vertices.

IV. CITATION DYNAMICS FROM THE PERSPECTIVE OF CITED PAPER

In this section we consider our scenario (Fig. 1) from the perspective of a cited paper rather than from the author's perspective and formulate the model of citation dynamics. This is equivalent to moving from the mean-field approach to the paper-specific approach. The model contains empirical functions that shall be taken from measurements. To find these functions we perform measurements of citation dynamics of a small set of physics papers and come to the conclusion that our linear model misses some important ingredient. We perform a series of microscopic measurements, pinpoint the missing ingredient (nonlinearity), and reformulate the model correspondingly.

A. Model: Citation dynamics of individual papers

To model citation dynamics of individual papers we reformulate our scenario (Fig. 1) in terms of citations. Figure 8 shows a source paper i published in the year t_0 and its citations garnered in subsequent years. A direct citation is the paper that cites paper i and does not cite any other paper that cites i, while an indirect citation is the paper that cites both i and one or more of its citing papers. For example, the papers A,B,C cite paper i and these are direct citations. The papers e,g,f cite, correspondingly, the papers B,f,C that cite paper i and these are indirect citations.

To quantify this scenario we designate by $k_i(t_0,t_0+t)$ the number of citations garnered by a paper i in the year t_0+t. Our basic assumption is that $k_i(t_0,t_0+t)$ is a discrete random variable that follows a time-inhomogeneous Poisson process [53] with the time-dependent Poissonian rate $\lambda_i(t_0,t_0+t)$ and the probability distribution

$$\mathrm{Poiss}(k_i)=\frac{\lambda_i^{k_i}}{k_i!}e^{-\lambda_i}. \tag{18}$$

Our aim is to model $\lambda_i(t_0,t_0+t)$. We assume that $\lambda_i(t_0,t_0+t)$ consists of the direct and indirect contributions,

$$\lambda_i(t_0,t_0+t)=\lambda_i^{\mathrm{dir}}(t_0,t_0+t)+\lambda_i^{\mathrm{indir}}(t_0,t_0+t). \tag{19}$$

We do not make any assumptions with respect to the functional form of $\lambda_i^{\mathrm{dir}}(t_0,t_0+t)$ but seek consistency with Sec. III B, namely, we require that

$$\overline{\lambda_i^{\mathrm{dir}}(t_0,t_0+t)}=M_{\mathrm{dir}}(t_0,t_0+t), \tag{20}$$

where $M_{\mathrm{dir}}(t_0,t_0+t)$ is given by Eq. (16) and the averaging is performed over all papers in one research field published in one year.

With respect to indirect citations, we assume that their dynamics is captured by the "individualized" version of Eq. (17), namely

$$\lambda_i^{\mathrm{indir}}(t_0,t_0+t)=\sum_{\tau=0}^{t}M(t_0+\tau,t_0+t)P(t_0+t,t_0+\tau)\times k_i(t_0,t_0+\tau), \tag{21}$$

where $k_i(t_0,t_0+\tau)$ is the number of previous citations of the source paper i garnered in the year $t_0+\tau$. We assume that functions M and P are the same for all papers in one field and published in one year (this will be revised soon). Under these assumptions Eq. (21) is consistent with Eq. (17) since $\overline{k_i(t_0,t_0+\tau)}=M(t_0,t_0+\tau)$ and thus $\overline{\lambda_i^{\mathrm{indir}}(t_0,t_0+t)}=M_{\mathrm{indir}}(t_0,t_0+t)$.

To cast Eq. (21) into a more concise form we substituted there exponential kernel $P(t_0+t,t_0+\tau)=P_0(t_0+t)e^{-\gamma(t-\tau)}$ found in our studies of references (Sec. II C 2). To reduce all functions to the same publication year t_0 we used the relations $P_0(t_0+t)=P_0(t_0)e^{-\beta t}$ and $M(t_0+\tau,t_0+t)=M(t_0,t_0+t-\tau)e^{\beta\tau}$ where β accounts for the growth of the reference list length with time (see Secs. II C 1 and II C 3). We exclude t_0 from our notation and come to

$$\lambda_i^{\mathrm{indir}}(t)=\sum_{\tau=0}^{t}M(t-\tau)P_0e^{-(\gamma+\beta)(t-\tau)}k_i(\tau). \tag{22}$$

Our purpose is to measure $\lambda_i^{\mathrm{dir}}(t)$ and $\lambda_i^{\mathrm{indir}}(t)$ for individual papers and to compare our measurements to Eqs. (19), (20), and (22), correspondingly.

B. Methodology: Citation dynamics of the groups of similar papers

Comparison of citation dynamics of a single paper to model prediction is not very instructive since this dynamics has too much variability for individual papers. To make meaningful comparison and to minimize scatter we considered average citation dynamics of the groups of similar papers. We assumed that the papers that belong to the same research field, were published in the same year, *and garnered the same number of citations in the long-time limit*—K_∞—have similar citation dynamics. In other words, we assumed that the paper's individuality is captured by K_∞. In particular, we measured citation dynamics of four groups of papers that garnered, correspondingly, 10, 20, 30, and 100 citations in the long-time limit and assumed that $\lambda_i(t)=\overline{k_i(t)}$ where $\overline{k_i(t)}$ is the mean citation rate of the papers with the same K_∞. To reduce noise,

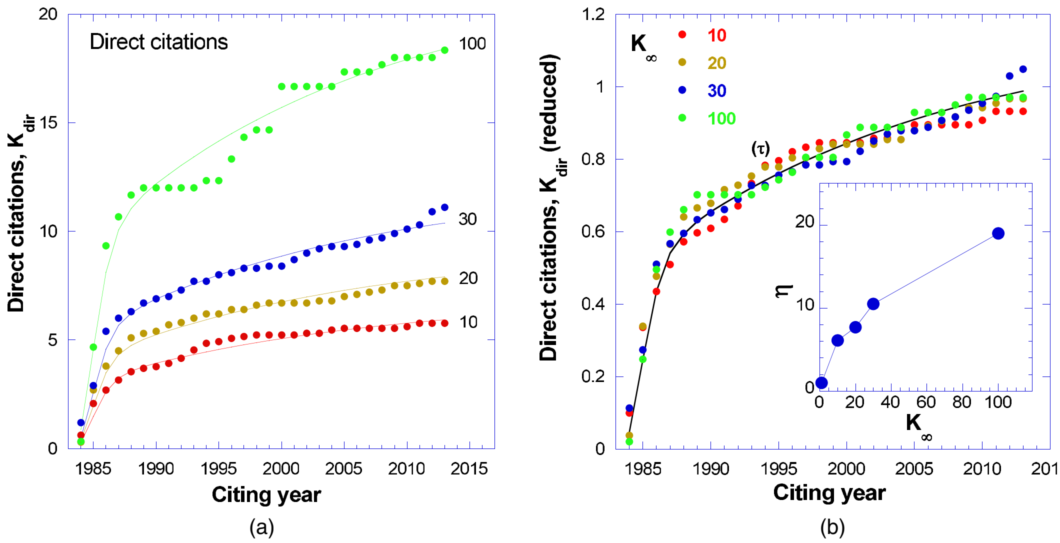

FIG. 9. (a) Direct citations for 37 *Physical Review B* papers published in 1984. Each set of points represents cumulative direct citations averaged over a group of papers that garnered the same number of citations by the end of 2013, namely $K_\infty = 10$, 20, 30, and 100. Continuous lines show $K_{\rm dir}(t) = \eta(K_\infty)\sum_{\tau=0}^{t} m_{\rm dir}(\tau)$ dependencies where $m_{\rm dir}(t)$ is the same function for all groups and $\eta(K_\infty)$ is the fitting parameter for each group. (b) Scaled data of (a). Continuous black line shows $\sum_{\tau=0}^{t} m_{\rm dir}(\tau)$ dependence which was obtained by averaging and smoothing the scaled data. The inset shows $\eta(K_\infty)$. The line there is a guide to the eye.

comparison to the model was performed using cumulative citations $K_i(t) = \sum_{\tau=0}^{t} k_i(\tau)$. Our aim is to find $\lambda_i(t)$ and to verify whether the function $M(t-\tau)$ and the parameters P_0 and γ are the same for all papers.

This approach is equivalent to replacing Eqs. (19) and (21) by Eqs. (15) and (17) where averaging is performed not over a whole set of papers in the field, but over a subset consisting of papers having the same number of citations in the long-time limit. Although this mesoscopic approach is much more specific than the mean-field one analyzed in Sec. III, still it captures only the deterministic component of citation dynamics while washing away its variability. The variable, stochastic component of citation dynamics, will be considered in Sec. V.

C. Direct citations

Figure 9 shows time dependence of $K_{\rm dir}(t) = \sum_{\tau=0}^{t} k_i^{\rm dir}(\tau)$, cumulative direct citations for several groups of papers that were published in the same year. $K_{\rm dir}(t)$ dependencies for all groups are qualitatively similar and after scaling they collapse onto a single curve. This means that they can be represented as $K_{\rm dir}(t) = \eta(K_\infty)\sum_{\tau=0}^{t} m_{\rm dir}(\tau)$ where K_∞ is the long-time limit of the number of citations, $\eta(K_\infty)$ is the scaling factor, and $m_{\rm dir}(t)$ is the universal function for all groups (see Fig. 10). Since $K_{\rm dir}(t)$ does not come to saturation even at $t = 30$ years, in order to uniquely define $m_{\rm dir}(t)$ we adopted the following constraint: $\sum_{\tau=0}^{t=30} m_{\rm dir}(\tau) = 1$. Under this constraint η_i is the number of direct citations that paper i garners after 30 years. We name it fitness. Then

$$\lambda_i^{\rm dir}(t) = \eta_i m_{\rm dir}(t). \tag{23}$$

Thus, the average annual number of direct citations for all papers in one field and published in one year is $\overline{k_i^{\rm dir}(t)} = \overline{\eta_i} m_{\rm dir}(t)$. According to our model [Eq. (16)] this expression shall be equal to $M_{\rm dir}(t) = r_{\rm dir}(t)R_0(t_0)e^{(\alpha+\beta)t}$ where $r(t)$ is the reduced reference age and R_0 is the average reference list length. Figure 9(c) validates this relation—the functions $m_{\rm dir}(t)$ and $r_{\rm dir}(t)$ are very similar and their ratio is close to expected exponential dependence.

D. Indirect citations

Figure 11 plots $K_{\rm indir}(t)$, cumulative indirect citations for the groups of papers shown in Fig. 9. In our calculation we

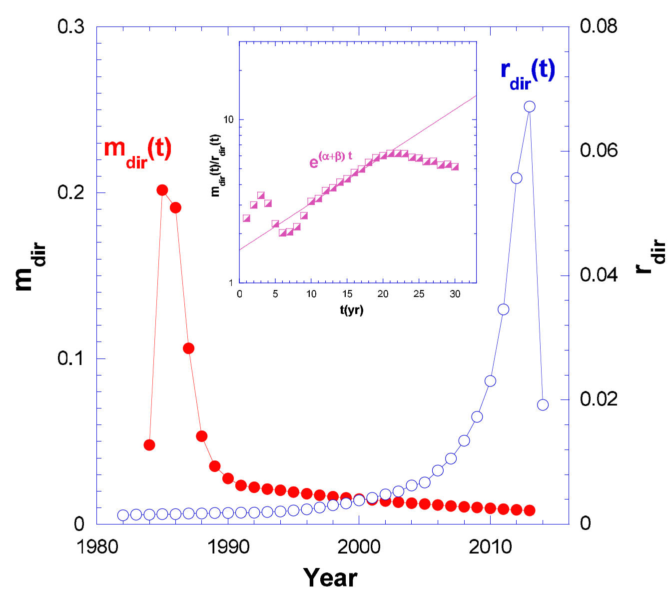

FIG. 10. Time dependence of reduced direct citations $m_{\rm dir}(t)$ [Eq. (23), Fig. 9(b)] and of reduced direct references $r_{\rm dir}(t)$ (from Fig. 4). The inset shows $m_{\rm dir}(t)$ to $r_{\rm dir}(t)$ ratio. The straight line indicates exponential dependence $e^{(\alpha+\beta)t}$ suggested by Eq. (14) with $\alpha = 0.046$ yr^{-1} and $\beta = 0.02$ yr^{-1} as found in SM III [39].

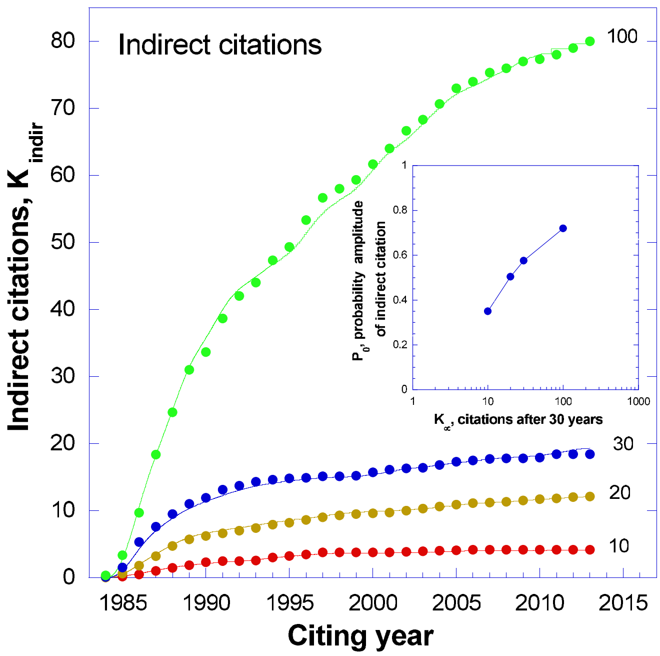

FIG. 11. Indirect citations for 37 *Physical Review B* papers published in 1984. Each set of points represents cumulative indirect citations averaged over a group of papers that garnered the same number of citations K_∞ (10, 20, 30, and 100) by the end of 2013. Continuous lines are fits to Eq. (22) with $\gamma = 1.2\,\mathrm{yr}^{-1}$, $\beta = 0.02\,\mathrm{yr}^{-1}$, $M(t-\tau)$ from Fig. 6 and P_0 as a fitting parameter for each group. The inset shows $P_0(K_\infty)$ dependence. The line there is a guide to the eye.

used Eq. (22) where instead of $k_i(\tau)$ we substituted $\overline{k_i(\tau)}$, the mean citation rate for the group of papers with the same K_∞. The function $M(t)$ was taken from Fig. 6 while P_0, γ, and β were taken from our measurements of references. (Note that P_0 depends on the publication year. Our studies of references of the physics papers published in 2014 yielded $P_0 = 0.19$. Extrapolation to 1984 based on the results of Sec. II C 3 yields $P_0 = 0.31$.) The model prediction satisfactorily fits $K_{\mathrm{indir}}(t)$ dependence for low-cited papers with $K_\infty = 10$ and is inconsistent with our measurements for the papers with $K_\infty > 10$. However, if P_0 is considered as a fitting parameter for each group, then a good agreement between the measured and calculated $K_{\mathrm{indir}}(t)$ dependencies is achieved (Fig. 11). The inset to Fig. 11 shows that the fitting parameter P_0 increases with K_∞. In fact, Eq. (21) implies that K_∞ dependence may be attributed either to P_0, or to $M(t-\tau)$, or to both of them. It is important to note that $M(t-\tau)$ in Eq. (22) is the mean number of citations of the papers that cite the source paper, i.e., it is the nearest-neighbor connectivity M^{nn} which is associated with the second-generation citing papers. The probability of indirect citations is also related to the second-generation citing papers (next-nearest neighbors). Therefore, to find the origin of the $P_0(K_\infty)$ dependence we decided to study the second-generation citations and citing papers more closely.

E. Second-generation citations and citing papers

We again considered several groups of physics papers published in one year (1984) and having the same number of citations in the long-time limit (2014). For each of these source papers we counted their first- and second-generation citing papers and citations garnered by 2014. Obviously, for each source paper the numbers of the first-generation citations and citing papers are equal. However, the number of the second-generation citations generally exceeds the number of the second-generation citing papers since one second-generation paper can cite several first-generation citing papers. We denote by M_i^{nn} and N_i^{nn}, correspondingly, the number of the second-generation *citations* and the number of the second-generation *citing papers* per one first-generation citing paper. In the language of network science M_i^{nn} is the average nearest-neighbor connectivity while N_i^{nn} is the average number of next-nearest neighbors per one nearest neighbor. Both M_i^{nn} and N_i^{nn} increase with time and for most of the papers these parameters achieve saturation in the long-time limit. We calculated $M^{nn}(K_\infty) = \overline{M_i^{nn}}$ and $N^{nn}(K_\infty) = \overline{N_i^{nn}}$ where the averaging was performed over a group of papers with the same K_∞, the number of citations in the long-time limit. Figure 12 shows that $M^{nn}(K_\infty)$ slowly increases, while $N^{nn}(K_\infty)$ is nearly independent of K_∞. Increasing $M^{nn}(K_\infty)$ dependence indicates that highly cited papers have highly cited descendants, i.e., citation network is assortative. Reference [2] made a similar observation for the network of *Physical Review* to *Physical Review* citations. It is important to note that M_i^{nn} and N_i^{nn} for the same paper are correlated and large N_i^{nn} usually means large M_i^{nn}. To account for this correlation we introduced a new parameter $s_i = \frac{M_i^{nn}}{N_i^{nn}}$ that characterizes the mean number of paths connecting the source paper to its next-nearest neighbors and which is closely related to the so-called quadrangle coefficient [54]. Figure 12(b) shows $s = \overline{s_i}$ where the averaging was performed over the groups of papers with the same K_∞. The error bars indicate the spread of s_i values within each group. This spread is much smaller than those of M_i^{nn} and N_i^{nn} [Fig. 12(a)], as expected.

Figure 12(b) suggests that s grows logarithmically with K_∞ and increases from $s \approx 1$ for low-cited papers to $s = 1.55$ for highly cited papers. This means that the former are connected to their second-generation descendants mostly by single paths, while the latter are connected to their second-generation descendants by multiple paths. The difference between the network neighborhoods of the lowly and highly cited papers may arise from the saturation effect: the descendants of lowly cited papers constitute only a small fraction of all papers in the corresponding research field, while the descendants of highly cited papers constitute a considerable fraction of it (see Appendix A).

F. Probability of indirect citation

Although $M^{nn}(K_\infty)$ dependence (assortativity) introduces some K_∞ dependence into the kernel of Eq. (22), this dependence is too weak and cannot qualitatively explain the K_∞-dependent factor (which we attributed to P_0) that was invoked in order to make our measurements of indirect citations consistent with Eq. (22) (see Fig. 11). This indicates that not only is $M(t-\tau)$ [which is in fact $M^{nn}(t-\tau)$] K_∞ dependent, but the factor P_0 should also depend on K_∞. In view of Fig. 12 we speculate that $P_0(K_\infty)$ dependence can be

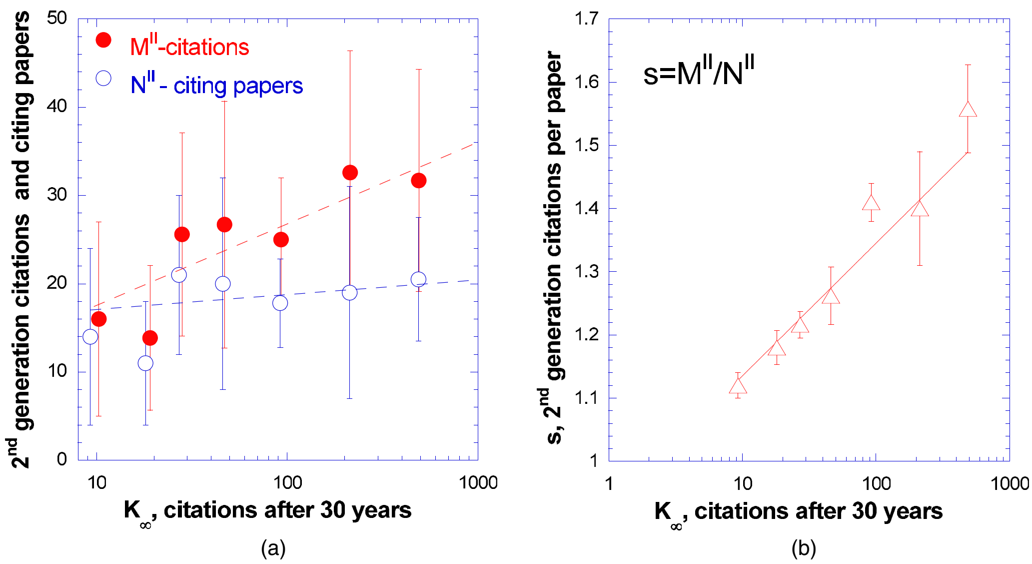

FIG. 12. (a) Second-generation citations and citing papers for 108 *Physical Review B* papers published in 1984. The filled circles show the average nearest-neighbor connectivity, $M^{nn}(K_\infty) = \overline{M_i^{nn}}$. The empty circles show the average number of next-nearest neighbors per one nearest neighbor, $N^{nn}(K_\infty) = \overline{N_i^{nn}}$. The averaging was performed over the groups of papers with the same K_∞, the number of citations garnered by the end of 2013. While $N^{nn}(K_\infty)$ is nearly independent of K_∞, $M^{nn}(K_\infty)$ increases with K_∞, indicating that citation network is assortative. The dashed lines are the guides to the eye. (b) The mean number of paths connecting a second-generation citing paper to the source paper, $s = \overline{s_i}$, where $s_i = \frac{M_i^{nn}}{N_i^{nn}}$. The straight line shows empirical logarithmic dependence $s = 0.925 + 0.21 \log_{10} K_\infty$.

traced to the fact that the network neighborhoods of the lowly and highly cited papers differ.

Figure 13 illustrates how this can occur. It shows some source paper i and its first- and second-generation descendants. The decision on whether to cite indirectly (to copy) the source

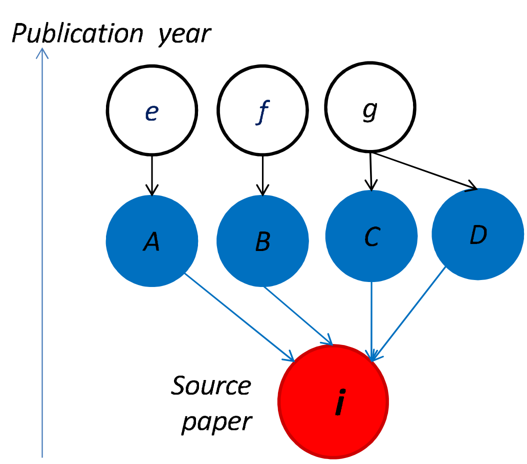

FIG. 13. Probability of indirect citation depends on the number of paths connecting the citing and cited papers. With respect to the source paper i the papers A,B,C and e,f,g are, correspondingly, the first-generation and second-generation citing papers. The probabilities that the papers e,f cite the source paper i are determined by the paths e-A-i and f-B-i. Since the papers e and f are written by different authors these paths do not interfere and the corresponding probabilities sum up. The paper g cites two first-generation citing papers C and D; the probability that it cites the source paper i is determined by the paths f-C-i and f-D-i. Since the decision on copying is made by the same author these paths interfere and the corresponding probabilities mix nonlinearly.

paper or not is made at the second-generation node rather than by the second-generation link. This prompts us to modify Eq. (22): we replace there $M_i^{nn}(t-\tau)$ by $N_i^{nn}(t-\tau)$. The probability P_0 is replaced, correspondingly, by $\tilde{P}_0 = \frac{M_i^{nn}}{N_i^{nn}} P_0 = s_i P_0$. While P_0 is the probability of copying the source paper which is induced by a second-generation citation, $\tilde{P}_0$ is the probability of copying the source paper by a second-generation citing paper. $\tilde{P}_0$ takes into account that one second-generation citing paper can cite several first-generation citing papers. After this replacement Eq. (22) reduces to

$$\lambda_i^{\text{indir}}(t) = \sum_{\tau=0}^{t} N^{nn}(t-\tau)\tilde{P}_0(K_\infty)e^{-(\gamma+\beta)(t-\tau)}k_i(\tau). \qquad (24)$$

Since N^{nn} is almost independent of K_∞ [Fig. 12(a)] then $\tilde{P}_0(K_\infty)$ absorbs all K_∞-dependent factors. Thus, the inset to Fig. 11 shows in fact the $\tilde{P}_0(K_\infty)$ dependence. To find the origin of this dependence we shall consider the copying mechanism at the microscopic level.

Our basic assumption is that $\tilde{P}_0$ is sensitive to the network neighborhood of the source paper. To study this issue in detail we chose three representative *Physical Review B* papers that were published in 1984 and gained 100 citations by the end of 2013. For each source paper we considered two generations of citing papers, limited ourselves only to descendants of the direct citations, and disregarded indirect citations inducing other indirect citations. We designated the number of direct citations of a source paper by K_{dir}. The number of papers that cite these K_{dir} papers is designated by N^{II} and the number of corresponding citations is M^{II} (all citations were counted by the end of 2013). Among these N^{II} papers we counted the number of network motifs (j-multiplets; see Fig. 14) which we designated by N_j^{II}. The fraction of the second-generation

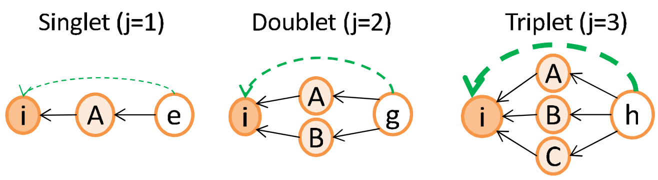

FIG. 14. Network motifs. The circles show papers, the continuous lines show direct citations, the dashed lines show indirect citations. i: the source paper; A,B,C: first-generation citing papers; e,g,h: second-generation citing papers. We distinguish between j-multiplet such as singlet ($j=1$), doublet ($j=2$), triplet, ($j=3$), etc. Figure 15 indicates that the probability of papers e,g,h to cite the source paper i (indirectly) increases nonlinearly with the multiplicity j.

citing papers associated with j-multiplets is f_j, in such a way that $N^{II}=\sum_j f_j N_j^{II}$ and $M^{II}=\sum_j jf_j N_j^{II}$. Among each group of N_j^{II} papers we counted those that cite the source paper (indirect citations) and designated their number as $(N_j^{II})_{\text{indir}}$. Thus, the probability of indirect citation of the source paper by a second-generation citing paper which is a part of a j-multiplet is $\pi_j=\frac{(N_j^{II})_{\text{indir}}}{N_j^{II}}$.

Figure 15 shows that $\pi_j(j)$ dependence is nonlinear and this is highly nontrivial. Indeed, if each second-generation citation were having the same probability of inducing indirect citation of the parent paper, this probability should increase linearly with the number of paths connecting the citing paper

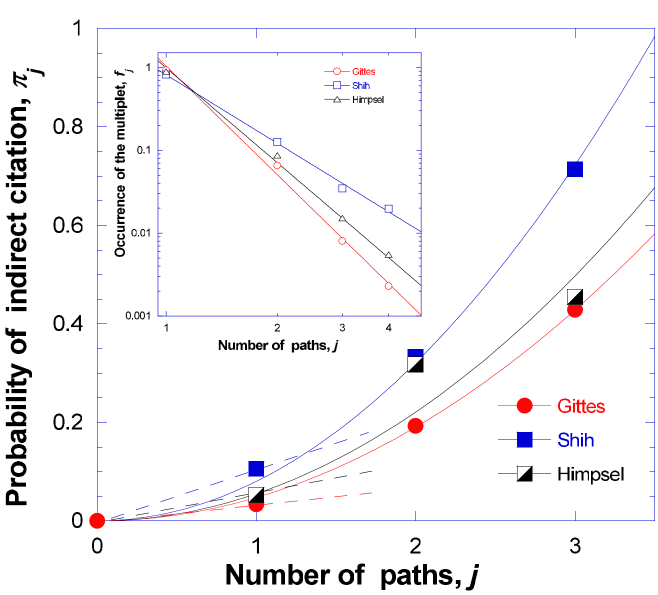

FIG. 15. Probability of indirect citation in network motifs vs j, the number of paths connecting the citing and cited paper. For each source paper we plot the linear ($\pi_j \propto j$, dashed line) and quadratic ($\pi_j \propto j^2$, continuous line) dependencies. Quadratic dependence, suggestive of interference, fits the data much better than the linear one. The inset shows that f_j, the fraction of network motifs (Fig. 14) among all second-generation citing papers, decreases with increasing multiplicity j as $f_j \sim j^{-d}$ where $d=3$–4. The data are for three *Physical Review B* papers that were published in 1984 and gained 100 citations by the end of 2013.

to its ancestor, namely, $\pi_j \propto j$. Figure 15 indicates that π_j rather follows quadratic dependence, $\pi_j \propto j^2$, suggestive of multipath interference. Thus, the probability of citing the source paper by any second-generation citing paper which is a part of a high-order multiplet (quadrangle, pentagon, etc.) is disproportionately high. Is this effect important? Although f_j, the fraction of higher-order j-multiplets, decreases with j, the contribution of these higher-order multiplets to the number of indirect citations is by no means negligible. Consider the papers shown in Fig. 15. While higher-order multiplets $j=2,3\ldots$ constitute only 12% of the second-generation citing papers they contribute 44% of indirect citations.

These microscopic measurements allow quantitative assessment of the $\tilde{P}_0(K)$ dependence. Indeed, we note that $\tilde{P}_0 \propto \sum_j \pi_j f_j$ where f_j is the fraction of the higher-order multiplets. For simplicity, we limit ourselves only to singlets and doublets, in such a way that $f_1+f_2=1$. Since $\pi \propto j^2$ (Fig. 15) then $\pi_2 \approx 4\pi_1$ and $\tilde{P}_0 \propto \pi_1(1+3f_2)$. We note that $s=\sum_j jf_j$ where s was introduced in Sec. IV E. Then $s=1+f_2$ and

$$\tilde{P}_0 \propto \pi_1[1+3(s-1)]. \tag{25}$$

If the multipath interference were absent, then $\pi_j \sim j$, in such a way that $\pi_2=2\pi_1$ and

$$\tilde{P}_0 \propto \pi_1 s. \tag{26}$$

Our microscopic measurements with 37 physics papers [Fig. 16(a)] are consistent with Eq. (25) rather than with Eq. (26) and for $K_\infty > 10$ they yield $\tilde{P}_0 \approx 0.42[1+3(s-1)]$. This is another proof of the multipath interference.

Since s depends on K_∞, the $\tilde{P}_0(s)$ dependence captured by Eq. (25) is translated into $\tilde{P}_0(K_\infty)$ dependence. Indeed, consider Fig. 16(b) which is the combination of Fig. 12(b) and the inset to Fig. 11. We scale the vertical axes as suggested by Eq. (25) and find that both $\tilde{P}_0(K_\infty)$ and $1+3(s-1)$ vs K_∞ dependencies collapse. This proves that the underlying cause for the $\tilde{P}_0(K_\infty)$ dependence is $s(K_\infty)$ dependence amplified by multipath interference.

Now we make a crucial assumption. Although Fig. 16(b) shows $\tilde{P}_0(K_\infty)$ dependence for $K_\infty = K(t=30)$, we note that there is nothing special about $t=30$ and the same dependence should hold for any year t. If we adopt this conjecture then Fig. 16(b) yields

$$\tilde{P}_0(K) = 0.34(1+0.82\log_{10} K), \tag{27}$$

where K is the number of citations of the paper at year t. Since $K(t)$ grows with time, so does $\tilde{P}_0(K)$. Thus we found an explanation of the puzzling $\tilde{P}_0(K)$ dependence shown in the inset of the Fig. 11. Now we can extend our citation model from the mesoscopic (groups of similar papers) to the microscopic (individual papers) level.

V. STOCHASTIC MODEL OF CITATION DYNAMICS

In the previous section we formulated the model of citation dynamics for individual papers. The model contained several empirical functions. To calibrate the model we considered several groups of similar papers, measured average citation dynamics of the papers in each group, fitted them using the model, and found the corresponding empirical functions.

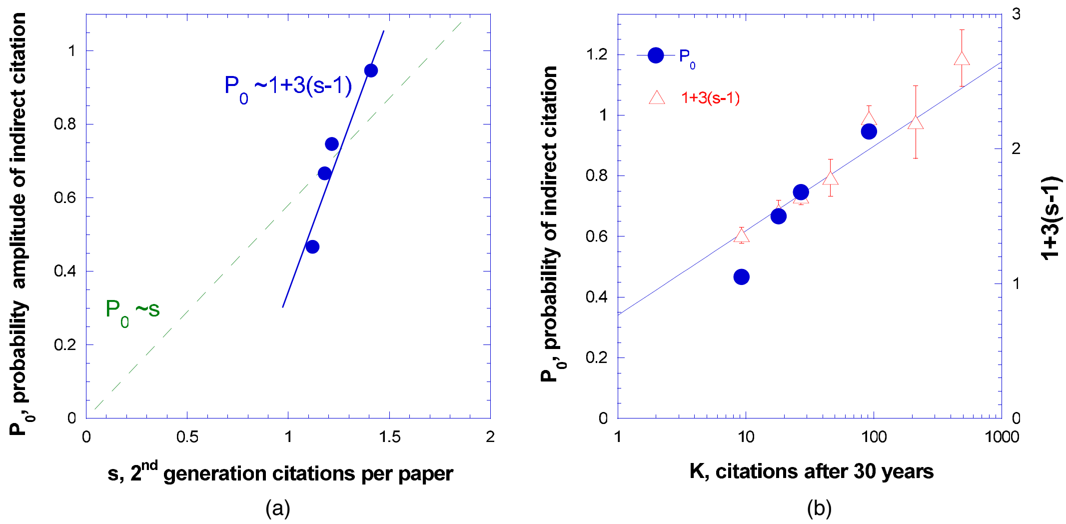

FIG. 16. (a) $\tilde{P}_0$, the probability of indirect citation of the source paper by a second-generation citing paper, vs s, the average number of paths connecting these two papers. The data are for 37 papers shown in Fig. 11. The straight blue line shows a fairly good fit to Eq. (25) with π_1 as a fitting parameter. The dashed green line shows fit to Eq. (26). Clearly, this fit is unsatisfactory. (b) $\tilde{P}_0$ vs K_∞, the number of citations garnered by a paper after 30 years (filled circles). Empty triangles show $1+3(s-1)$ where s values were taken from Fig. 12. The straight line shows logarithmic dependence given by Eq. (27).

The calibrated model captures the deterministic component of citation dynamics of papers. However, it is not clear whether it reproduces its random, fluctuating component.

The aim of this section is the validation of the model. This shall be done for two reasons. First, since we measured empirical functions appearing in the model by studying citation dynamics of the papers that garnered only 10–100 citations in the long-time limit we need to verify that our model can be extrapolated to all papers, namely, to those that garnered many more and many fewer citations. Second, we need to check to what extent our model captures stochastic component of citation dynamics of papers.

A. Model formulation

For pedagogical reasons we summarize here our model [Eqs. (23), (24), and (27)] and present our key result—the nonlinear stochastic dynamic equation for the latent citation rate of a paper i at year t after publication,

$$\lambda_i(t) = \eta_i m_{\text{dir}}(t) + \sum_{\tau=0}^{t} N^{nn}(t-\tau)\tilde{P}_0(K_i)e^{-(\gamma+\beta)(t-\tau)}k_i(\tau). \tag{28}$$

Here, η_i is the paper's fitness, an empirical parameter, unique for each paper; $m_{\text{dir}}(t)$ is the time-dependent direct citation rate; $k_i(\tau)$ is the actual number of citations that the paper i garnered earlier in the year τ; $N^{nn}(t-\tau)$ is the average number of the second-generation citing papers (per one first-generation citing paper) published in the year $t-\tau$; $\tilde{P}_0(K_i)e^{-(\gamma+\beta)(t-\tau)}$ is the probability of indirect citation of the paper i by a second-generation citing paper published in the year $t-\tau$; γ is the obsolescence exponent, and β is the exponent characterizing the growth of the reference list length with time. $k_i(t)$ is given by the Poisson distribution, $\text{Poiss}(k_i) = \frac{\lambda_i^{k_i}}{k_i!}e^{-\lambda_i}$. The exponents γ and β, the functions $m_{\text{dir}}(t)$, $N^{nn}(t-\tau)$, and $\tilde{P}_0(K)$ are the same for all papers in one field published in one year.

Equation (28) relates $\lambda_i(t)$, the latent citation rate of a paper, to its past citation rate $k_i(\tau)$ and to the number of previous citations $K_i(\tau)$ at all previous years, in other words, Eq. (28) describes a non-Markovian process with memory [55]. Viewing it from a different perspective we notice that Eq. (28) describes a self-exciting Hawkes process. Similar equations appear in the renewal theory, in the context of Bellman-Harris branching (cascade) processes [56], in population dynamics (the age-dependent birth-death process with immigration [57]), dynamics of viewing behavior of YouTube users [58], social networking sites (resharing) [59], and viral information spreading [60]. In distinction to these well-known cases, Eq. (28) is nonlinear, the nonlinearity arising from the $\tilde{P}_0(K)$ dependence.

B. Model validation

1. Methodology

If citation dynamics of individual papers were following a homogeneous stochastic process we could compare it to the model prediction on the paper-by-paper basis. However, citation dynamics of scientific papers is an inhomogeneous stochastic process which cannot be decomposed into mean and random parts. To perform a meaningful comparison to the model we adopted the following strategy. We considered a large ensemble of papers in the same field that were published in the same year (40 195 physics papers published in 83 leading physics journals in 1984) and measured citation dynamics of all these papers using Thomson-Reuters Web of Science database. In the framework of our model, the papers in this set differ only by their fitness. Then we designed a synthetic set of papers with the same number of papers and the same fitness distribution, performed numerical simulations of their citation dynamics, and compared them to model prediction.

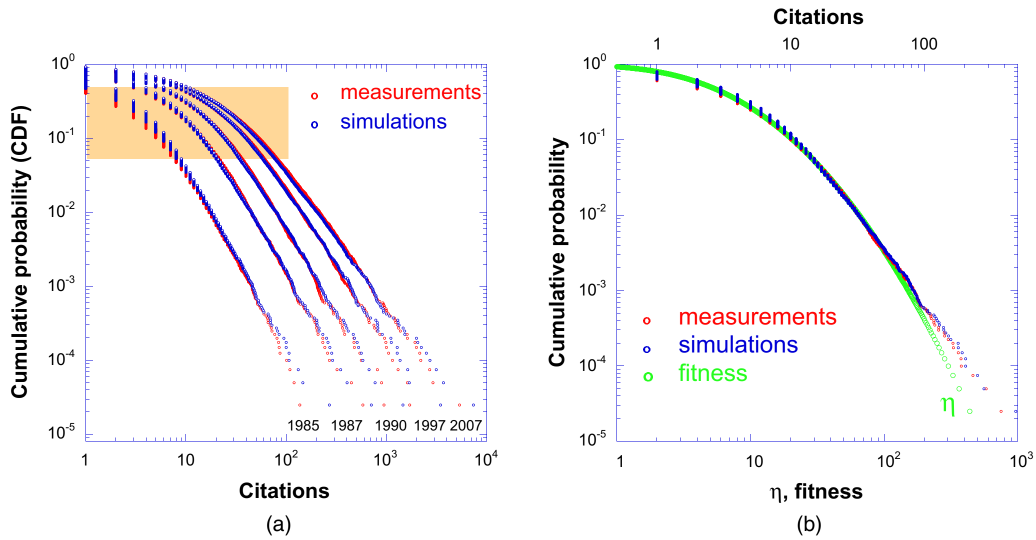

FIG. 17. (a) Annual cumulative citation distributions for 40 195 physics papers published in 1984. Red circles show measured data, blue circles show results of stochastic simulation based on the Poisson process with the rate given by Eq. (28). Hatched area shows the subset of papers with the same range of $K_\infty = 10$–100 as the sets of papers that were used in Sec. IV C to calibrate our model. (b) The set of fitnesses used in our simulation is described by the log-normal distribution with $\mu = 1.62$ and $\sigma = 1.1$ (green circles). The red and blue circles show the measured and simulated citation distributions for the year 1986, i.e., two to three years after publication. The fitness distribution mimics these distributions and deviates from them only for highly cited papers.

How to organize such comparison is by no means obvious. While earlier models of complex networks growth were validated mostly by comparing measured and simulated aggregate characteristics, such as degree distribution, Eq. (28) is the next-generation model which is much more detailed and the comparison to measurements is more demanding. To the best of our knowledge, the methodology of comparing the stochastic model or simulation to stochastic data is not well established. Following Ref. [61] we believe that the proper validation of a stochastic model shall include multidimensional analysis. In particular, we validated our model in several dimensions:

(1) cumulative citation distributions;
(2) stochastic component of the citation dynamics;
(3) citation trajectories of individual papers;
(4) autocorrelation of citation trajectories;
(5) the number of uncited papers.

In what follows we address items (1) and (3) and consider the rest in the Supplemental Material (SM IX [39]).

2. *Citation distributions*

Figure 17 shows measured citation distributions for the set of 40 195 physics papers published in 1984. To simulate these distributions we need to find the corresponding fitnesses. In the framework of our model, the probabilistic estimate of the paper's fitness η_i can be done based on the expression $\eta_i \approx \frac{K_i^{\rm dir}(t)}{\sum_{\tau=0}^{t} m_i(\tau)}$ where $K_i^{\rm dir}(t)$ is the number of direct citations of the paper at some moment t [see Eq. (28)]. To measure $K_i^{\rm dir}(t)$ for each paper in our set proved to be too time-consuming, hence we found the fitness distribution indirectly. Namely, we run the simulation [Eq. (28)] for 40 195 papers assuming a log-normal fitness distribution, $\frac{1}{\sqrt{2\pi}\sigma\eta_i} \exp -(\frac{(\ln \eta_i - \mu)^2}{2\sigma^2})$ where μ and σ were fitting parameters. Our aim was to achieve the best correspondence to a measured citation distribution at $t = 1986$. Figure 17 shows that the thus found fitness distribution mimics the early citation distribution for all but the highly cited papers. This is not unexpected since citations garnered during the first two to three years after publication are mostly direct.

The other parameters of the simulation were as follows. We used $\gamma + \beta = 1.2$ yr^{-1}, as found in our measurements of indirect references and citations; $m_{\rm dir}(t)$ from Fig. 10, and $\tilde{P}_0(K)$ from Eq. (27). We assumed that $N^{nn}(t)$ dependence mimics $M(t)$, namely $N^{nn}(t) = \frac{M(t)}{\bar{s}}$ where $\bar{s} = 1.2$ is the average over all physics papers published in 1984 (see Sec. IV E) and $M(t)$ is shown in Fig. 6.

Figure 17(a) shows excellent agreement between the measured and simulated citation distributions at all years. Moreover, it shows that our model, which was calibrated using the papers with $K_\infty = 10$–100 citations, can indeed be extrapolated to the papers having more or fewer citations.

3. *Citation trajectories*

At the next step we compared the measured and simulated citation trajectories. It should be noted that citation dynamics of papers follows a self-exciting Hawkes process which amplifies past fluctuations. Therefore, even for the same initial conditions, the spread of simulated citation trajectories is so wide that the comparison of measured and simulated trajectories on the paper-by-paper basis is meaningless. Therefore, we compared citation trajectories for the sets of similar papers.

Figures 18 and 19 show citation trajectories for the sets of papers that garnered the same number of citations in the long-time limit. If we perform averaging for each set, the measured and simulated citation trajectories agree well. This is not unexpected since the empirical function

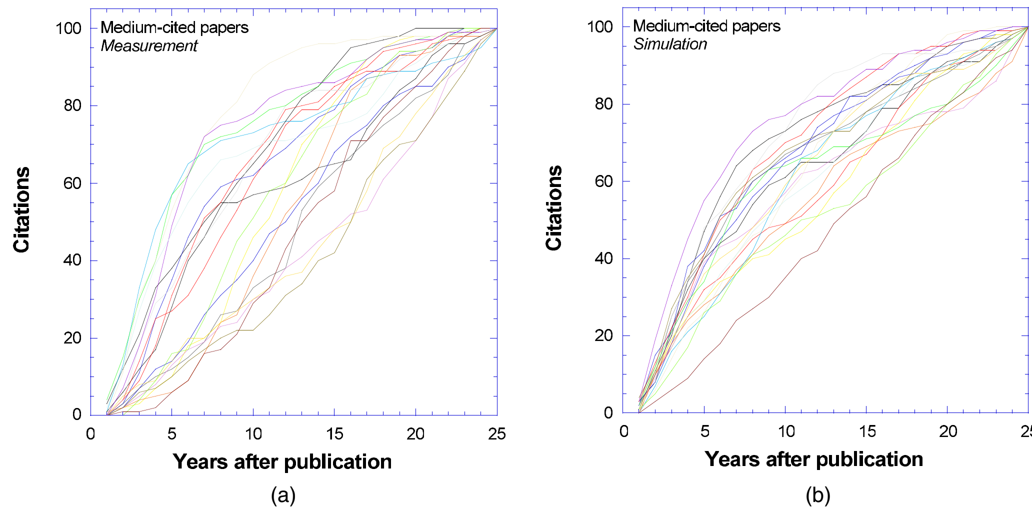

FIG. 18. Citation dynamics of the physics papers that were published in 1984 and accrued 99 citations in subsequent 25 years. Stochastic numerical simulation based on our model correctly predicts the shape and the spread of citation trajectories.

$\tilde{P}_0(K_i)e^{-(\gamma+\beta)(t-\tau)}$ [Eq. (28)] was established from the requirement that the model fits the average citation dynamics of similar papers. Thus, comparison of the shapes of the measured and simulated citation trajectories of individual papers in each group tells an independent story since the model does not contain free parameters to tailor these trajectories.

Figure 18 shows that for moderately cited papers the measured and simulated trajectories look very similar—they are jerky, and the fluctuations are of the same size. Figure 19 shows that for highly cited papers both sets of trajectories are smooth, but the spread of the measured trajectories exceeds that of the simulated ones.

4. Short summary

We found that Eq. (28) with log-normal fitness distribution reproduces citation dynamics of the physics papers fairly well. This includes aggregate characteristics (citation distributions) and microscopic dynamics (the number of uncited papers, the mean and the fluctuating parts of citation trajectories of individual papers, citation lifetime, etc.; see SM IX [39]). While our model correctly reproduces citation trajectories of the lowly and moderately cited papers, it underestimates the variability of citation trajectories of the highly cited papers.

VI. DISCUSSION

A. Comparison to previous studies

The closest predecessor of our model is the Simkin-Roychowdhury mathematical theory of citing [49] which is based on the copying algorithm of Krapivsky and Redner [20]. This theory is based on the following scenario: when a scientist writes a manuscript, he picks several recent papers published in the preceding year, cites them, and copies some

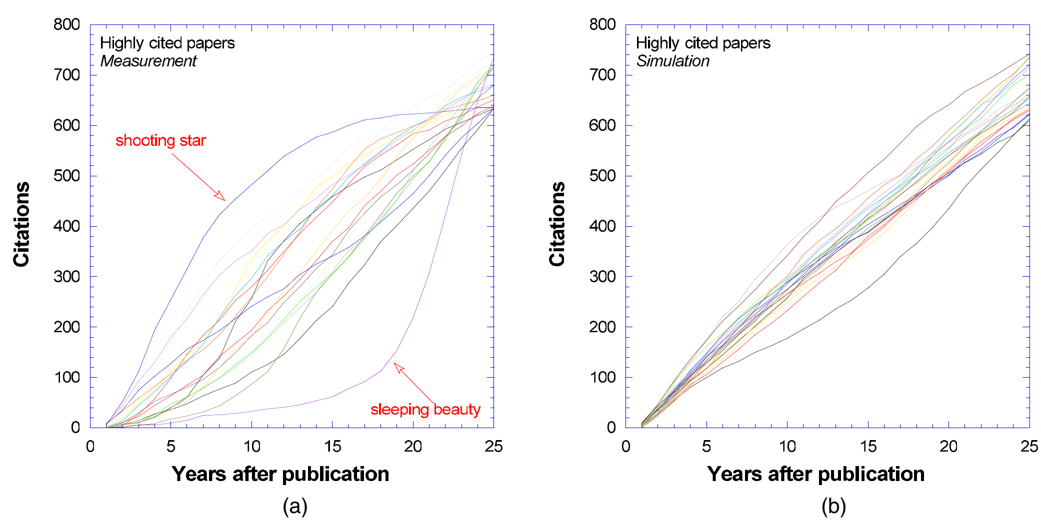

FIG. 19. Citation dynamics of the physics papers that were published in 1984 and accrued 600–750 citations in subsequent 25 years. For most of the papers the model correctly predicts the shape and smoothness of citation trajectories. However, the model does not capture extreme cases such as a sleeping beauty—the paper with delayed recognition—or a shooting star—the paper that is highly popular at the beginning of its citation career but then dies quickly.

of their references with equal probability. Thus, this model suggests a quickly decaying obsolescence function for direct pickup and age-independent probability of copying. This is the first-generation qualitative model—it provides clever insight, the basic scenario, but it cannot be used for qualitative estimates since the aging functions are speculative rather than taken from measurements. Our model is based on a much more detailed scenario: when a scientist writes a manuscript, he picks several papers with the probability depending on their publication year, cites them, and copies some of their references with the probability depending on the publication year of the parent paper and on the local structure of citation network associated with the copied paper. These probabilities are not speculative but are taken from measurements. Thus, our well-calibrated model builds on Refs. [20,49] but belongs to the next generation of models, those that can be used for quantitative estimates. Moreover, our model contains an additional ingredient—dynamic nonlinearity.

B. Nonlinearity and "power-law" degree distributions

Citations of scientific papers were one of the first examples of the power-law (fat-tailed) distribution [4]. The prevailing notion then was that all papers are created equal and the power-law distribution of their citations is created dynamically. Our results tell a different story: the fat-tail citation distribution is mostly inherited. Indeed, Fig. 17 shows that citation distribution at small t mimics fitness distribution which is already a fat-tailed distribution. As time goes on, citation distribution shifts towards higher K. Since the kernel in Eq. (28) is nonlinear and increases with K (the latter grows with time for each paper) the tail of the distribution shifts faster than its body. If the initial citation distribution was concave in the log-log coordinates, it straightens with time and approaches the power-law distribution. The ever decreasing slope of the fat-tail of citation distributions shown in Fig. 17 is a precursor of this transition from the convex to concave shape. This observation beats the intuition assuming that the power-law degree distribution is an evidence of the scale-free network. We show here that at least for citations, the power-law distribution is not the ultimate but a transient distribution.

Another consequence of nonlinearity is the appearance of runaways or "immortal papers" with infinite citation lifetime. As we already mentioned above, citation distribution shifts with time towards higher K whereas the tail of the distribution shifts faster than its body. Due to obsolescence, the body of the distribution eventually comes to stop but the tail may continue to shift. Thus, the papers in the tail exhibit "runaway" behavior—their citation career does not come to saturation even after a long time.

C. Preferential attachment

At the beginning of this research we believed that the citation network grows following Eqs. (2)–(4), which capture the preferential attachment rule. Hence, we based our model on a recursive search which is a specific implementation of the preferential attachment. Our measurements yielded Eq. (28) which is very different from the classical preferential attachment and rather follows the line of thought of Refs. [49,62] who focused entirely on fitness. Do our results invalidate the common understanding of the preferential attachment as an algorithm according to which a new node performs a global search in the whole network to find the most connected nodes? Not at all. In fact, our results suggest that the preferential attachment mechanism is indeed involved in citation network growth but it operates in a more subtle way than it was commonly believed.

Our measurements and modeling suggest the following mechanism of citation network growth. A new node in the network attaches to several older target nodes that become its nearest neighbors. Then the new node explores its next-nearest neighbors and preferentially connects to those of them that are linked to it through several nearest neighbors rather than through one of them. This procedure is similar to acquaintance immunization [63] and it finds the most connected nodes in the vicinity of the source node. Although this algorithm is based on the local search, it is one step towards the global search, since it analyzes not only the nearest, but also the next-nearest neighbors of the source node. Hence, the preferential attachment mechanism pops out explicitly in our model but in a different guise—it is captured by the kernel $\tilde{P}_0(K)$ in Eq. (28). Taken together with the assortativity of citation network, this algorithm results in the nonlinear attachment probability.

D. Prediction of citation trajectory of papers

The models of citation dynamics find application in predicting future citation trajectories of papers [10,64,65] and citation career of the authors [66–68]. Our calibrated model can be used for probabilistic prediction of the number of citations that a regular paper can garner in the future. Our formalism can be also used to pinpoint sleeping beauties or shooting stars at the earliest stage of their citation career. This task is usually solved by applying some model that extrapolates citation dynamics of papers from their citation history and then focuses on those papers that deviate from model prediction [61,69,70]. Our model is well suited for this purpose since it predicts not only the mean citation dynamic of a paper but the probability of its deviation from the mean as well. We leave for further studies application of our model for forecasting citation behavior of scientific papers.

What are the limitations of our model? One particularly strong assumption is the constancy of fitness along the whole citation career of the paper. Reference [11] in its description of the web pages popularity also used this assumption and justified it by measurements. While the assumption of constant fitness is reasonable for the majority of scientific papers and is validated by our measurements, there are sleeping beauties [Fig. 19(a)] that can be dormant for a long time and then achieve a burst of popularity. Although these papers are rare, they are often associated with scientific breakthroughs, and their importance is incomparable to their abundance. References [45,71] analyzed such papers and found that their peculiar citation trajectory (burst) has content-based explanation [65]. Is it possible that such citation bursts can appear by chance? Although our model describes a Hawkes process where small deviations from the mean-field behavior can be amplified during prolonged time period and thus produce bursts [72], we do not believe that our model can generate strong bursts. The reason is exponentially strong obsolescence γ [Eq. (28)] that prevents continuous amplification of small fluctuations. Thus we believe that our model describes a regular science in the sense of Kuhn [73] and does not capture exceptional papers associated with serendipitous discoveries, bursts of scientific activity, emergence of new and disappearance of old fields—everything that makes the science fun.

Equation (28) can be used for the prediction of citation rate of a regular paper. To do this we need to know the functions $m_{\rm dir}(t)$, $N^{nn}(t)$, $\tilde{P}_0(K)$, parameters γ,β (which are not paper-specific but depend on the research field and on the publication year), the past citation history of the paper, $k_i(\tau)$, and most important of all, its fitness. The function $\tilde{P}_0(K)$ and the exponent γ do not probably change with the publication year and can be estimated using citation histories of old papers in the field; the functions $N^{nn}(t)$ and $m_{\rm dir}(t)$ can be estimated on the basis of Eqs. (13) and (16), correspondingly; the exponents α and β, characterizing the annual growth of the number of publications and of the average reference list length, can be measured in the past and extrapolated to the future. The most tricky task is fitness estimation. Although the paper's fitness η_i shall be measured *a posteriori* when its citation career is ripe, some estimate of the paper's fitness can be done *a priori* on the basis of the number of direct or total citations garnered by a paper during the first two to three years after publication, which is nothing else but the paper's impact factor.

E. Comparison to other research fields

How general is our model? While it was calibrated using physics papers published in 1984, we performed similar measurements for mathematics and economics papers also published in 1984 and found very similar citation dynamics, including the nonlinear kernel. Hence we have a good reason to believe that our model describes these fields as well, albeit with different parameters. Namely:

(1) We found lognormal fitness distributions for physics, mathematics, and economics papers. The parameter μ, which characterizes the mean of the distribution, was different for these three fields. This obviously results from different citation practices: the reference list of a typical math paper is considerably shorter than that of a physics paper. Surprisingly, the parameter σ, characterizing the width of the distribution, was almost the same for all three fields.

(2) Indirect citations. Nonlinear kernel $\tilde{P}_0(K)$ with logarithmic dependence on the number of citations was found for all three fields.

(3) While we did not measure $m_{\rm dir}(t)$ for mathematics and economics, we expect that it is not the same as that for physics. $m_{\rm dir}(t)$ is determined by $r(t)$, age composition of the references in the references list, and α and β by the growth rates of the number of publications and the reference list length. While the $r(t)$ function seems to be very similar for different fields, the exponents α and β do differ.

F. Extension to other networks

We consider now a more general question—whether our network growth model, which is based on a recursive search with a nonlinear kernel, can describe other phenomena besides citations of scientific papers. Indeed, the mechanisms identical to a recursive search were invoked to account for spreading of ideas, rumors, diseases, and viral marketing [28,30,74–76]. Generally, these processes are modeled using linear dynamic equations assuming pairwise interactions between the neighbors in the network. The studies of Centola [77,78] of the spreading in social networks revealed complex propagation with social reinforcement where simultaneous exposure to several active neighbors in the network is important. Such synergistic effects in propagation on networks were also considered theoretically [79,80] and found experimentally in epidemiology [81], where susceptibility of a person to infection may depend on the number of infected neighbors. Our studies suggest that such multiple-node interactions result in nonlinear dynamics of complex propagation in networks. Indeed, Ref. [60] found nonlinearity in the dynamics of viral marketing. (Namely, it observed a correlation between transmittivity and fan-out coefficient which is very similar to our observation of the correlation between the number of second-generation citations and the probability of indirect citation; see Fig. 16. We showed that this correlation results in nonlinear dynamics.) References [82–84] found nonlinearity in citation dynamics of U.S. patents; Ref. [54] found nonlinearity in their studies of the Internet connectivity and growth.

VII. SUMMARY

We report a nonlinear stochastic model of citation dynamics of scientific papers. The model is fully calibrated by measurements of citations dynamics and statistics of references of physics papers. The model assumes that the author of a new scientific paper finds relevant papers from the media or journals and cites them. Then he studies the reference lists of these preselected papers, picks up relevant papers, cites them as well, and continues this process recursively. If some paper is cited by several preselected papers, the author chooses it with higher probability than those cited by only one preselected paper. This local rule enables the author to sample the global connectivity of the network.

This recursive search algorithm results in dynamic nonlinearity which is the reason why the ideas advocated in highly cited papers undergo viral propagation in the scientific community, while the low-cited papers affect only a small part of it. Such dynamic nonlinearity can play an important role in viral propagation in social media.

ACKNOWLEDGMENTS

We are grateful to S. Redner, L. Muchnik, D. Shapiro, M. Pollak, Y. Louzoun, and H. Brot for fruitful discussions. We appreciate instructive correspondence with M. Simkin. We are grateful to A. Scharnhorst for constant encouragement, and we acknowledge financial support of the EU COST Action TD1210. We are indebted to anonymous reviewers for their thorough and instructive suggestions that made our paper much more readable.

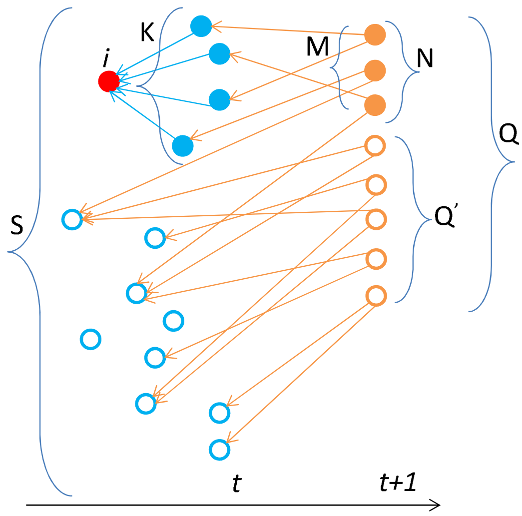

FIG. 20. A fragment of citation network showing a parent paper i and its first- and second-generation citations. There are K first-generation citing papers published during the period of t years after the publication of the source paper. These represent a subset of all S papers published in this field by year t. There are Q papers in this field, published in the year $t+1$, which cite some of S papers published earlier. Among these Q papers there are N second-generation citing papers that cite one of the K first-generation citing papers, and there are Q' papers that do not cite them. M is the number of the second-generation citations of the paper i published in the year $t+1$.

APPENDIX A: A HAND-WAVING EXPLANATION OF THE NONLINEAR PROBABILITY OF INDIRECT CITATION

We consider one possible source of nonlinear citation dynamics arising from the fact that $\tilde{P}_0(K)$, the probability of indirect citations of a paper, depends on the number of its previous citations K. We found that at the core of nonlinearity is the assortativity of citation network. We present here a toy model explaining this assortativity. Consider a parent paper i that has K citing papers published by year t (see Fig. 20). These K first-generation citing papers constitute a small part of a large set of all S papers that were published in this research field by year t. We denote by Q the total number of papers in this field that were published in the year $t+1$. We neglect obsolescence and assume that each of these Q papers issues on average $\sim m$ citations to the papers published previously. The total number of *citations* of all first-generation citing papers is $M \approx mQ\frac{K}{S}$. With respect to the parent paper i these are second-generation citations. The number of the corresponding second-generation *citing papers* is $N = Q - Q'$ where Q' is the number of papers published in the year $t+1$ that do not cite our K papers. (These definitions of M and N differ from those in the main text.) Assuming Poissonian distribution of citations issued by each paper from the Q set, we find $Q' = Q\sum_{n=0}^{\infty}(1-\frac{K}{S})^n\frac{m^n}{n!}e^{-m} = Qe^{-mK/S}\sum_{n=0}^{\infty}\frac{[m(1-\frac{K}{S})]^n}{n!}e^{-m(1-K/S)}$. According to the properties of the Poisson distribution, $\sum_{n=0}^{\infty}\frac{[m(1-\frac{K}{S})]^n}{n!}e^{-m(1-K/S)}=1$, hence $N = Q(1-e^{-mK/S})$.

We consider now the parameter s which is the average number of paths connecting a second-generation citing paper to the source paper i, namely, $s = \frac{M}{N} = \frac{m\frac{K}{S}}{1-e^{-m(K/S)}}$. We assume that $m\frac{K}{S} \ll 1$. We perform the series expansion of the above expression in small parameter $m\frac{K}{S}$ and retain the leading term in K: $s \approx 1 + K\frac{m}{2S}$. Thus s increases with K and this means that the highly cited papers have an increased proportion of multiple paths than the lowly cited papers. The source of nonlinear citation dynamics is this $s(K)$ dependence.

Of course, this hand-waving explanation of the $s(K)$ dependence does not account for all our results. It assumes that the number of second-generation citations of a given paper grows linearly with K while the number of its second-generation citing papers grows more slowly than linear with K. Our measurements indicate exactly the opposite behavior—the number of second-generating citing papers grows linearly with K and the number of second-generation citations grows faster than linear. Thus this toy model serves for purely illustrative purposes and cannot be used for calculations.

APPENDIX B: OUR RESULTS IN THE CONTEXT OF NETWORK SCIENCE

We consider our measurements of the direct and indirect citations in the context of network science. On the one hand, the number of second-generation citations M^{II} is nothing else but the average nearest-neighbor connectivity k_{nn}. Increasing $M^{II}(K)$ dependence indicates assortativity of citation network. On another hand, the number of indirect citations is related to the local clustering coefficient C_K, which is the ratio of the number of transitive triples to the total number of triples connected to a certain parent node. Indeed, consider a parent paper i that garnered K citations. The number of all triples connected to this paper is $N^{II}K$ where N^{II} is the average number of citing papers per one first-generation citing paper. Among these $N^{II}K$ papers there are some associated with indirect citations that participate in j-multiplets (Fig. 17). The number of the latter is $jf_j\pi_jN^{II}K$, where f_j is the fraction of j-multiplets among second-generation citing papers, π_j is the probability of indirect citation, and the factor j in the sum appears because each indirect citation in the j-multiplet is associated with j triangles. The number of all possible

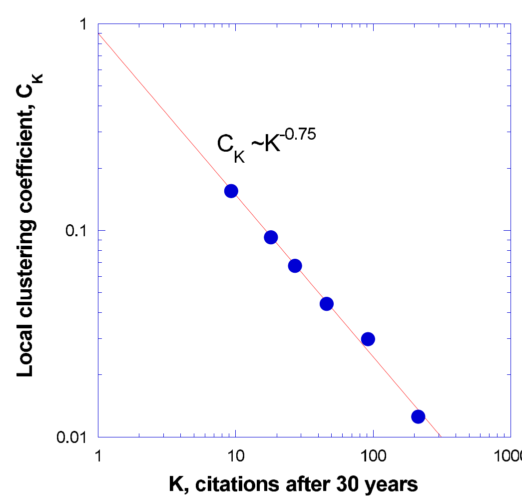

FIG. 21. C_k, local clustering coefficient. The filled circles show C_K calculated using Eq. (B2) and s from the Fig. 12(b). The straight line shows the power-law approximation.

triangles associated with the parent paper is $K(K-1)/2$. Then

$$C_K = \frac{2N^{II}\sum_{j=1} j\pi_j f_j}{K-1}. \quad \text{(B1)}$$

If we limit ourselves only to singlets and doublets and neglect higher-order multiplets, then $f_1 + f_2 \approx 1$ and $s \approx 1 + f_2$, where s is the ratio of the second generation citations to the second-generation citing papers, $s = \frac{M^{II}}{N^{II}}$. Our measurements suggest multipath interference, namely $\pi_2 = 4\pi_1$. Thus,

$$C_K \approx \frac{2N^{II}\pi_1[1+7(s-1)]}{K-1}. \quad \text{(B2)}$$

Our measurements indicate that N^{II} is almost independent of K. If s were independent of K, we expect that $C(K) \propto K^{-1}$. Our measurements show that s increases logarithmically with K [Fig. 12(b)]. Figure 21 shows that C_K, which was calculated according to Eq. (B2) using the data of Fig. 12(b), follows $K^{-0.75}$ dependence. This power-law dependence agrees with the findings of Ref. [2] for PR (*Physical Review*) to PR citation network.

Equation (B2) suggests an alternative interpretation of the probability of indirect citation $\tilde{P}_0$. Indeed, in Sec. IV F we showed that $\tilde{P}_0$ is determined by s, namely, $\tilde{P}_0 \propto \pi_1 [1 + 3(s-1)]$. By excluding $s-1$ from this equation and Eq. (B2) we find that $\tilde{P}_0 \propto C_K(K-1) + \text{const}$. This relation indicates that among the papers with the same number of previous citations, those with high clustering coefficient are cited more intensively—the possibility already considered theoretically by Bagrow and Brockmann [85].

[1] M. Newman, *Networks: An Introduction* (Oxford University Press, New York, 2010).
[2] A.-L. Barabasi, *Network Science* (Cambridge University Press, Cambridge, England, 2016).
[3] P. Holme and J. Saramaki, Phys. Rep. **519**, 97 (2012).
[4] D. D. S. Price, J. Am. Soc. Inf. Sci. **27**, 292 (1976).
[5] R. Albert and A.-L. Barabasi, Rev. Mod. Phys. **74**, 47 (2002).
[6] M. Perc, J. R. Soc. Interface **11** (2014).
[7] S. N. Dorogovtsev and J. F. F. Mendes, Phys. Rev. E **62**, 1842 (2000).
[8] P. L. Krapivsky and S. Redner, Phys. Rev. E **63**, 066123 (2001).
[9] G. Bianconi and A.-L. Barabasi, Phys. Rev. Lett. **86**, 5632 (2001).
[10] D. Wang, C. Song, and A.-L. Barabasi, Science **342**, 127 (2013).
[11] J. S. Kong, N. Sarshar, and V. P. Roychowdhury, Proc. Natl. Acad. Sci. USA **105**, 13724 (2008).
[12] M. Medo, G. Cimini, and S. Gualdi, Phys. Rev. Lett. **107**, 238701 (2011).
[13] F. Menczer, Proc. Natl. Acad. Sci. USA **101**, 5261 (2004).
[14] F. Papadopoulos, M. Kitsak, M. Á. Serrano, M. Boguñá, and D. Krioukov, Nature (London) **489**, 537 (2012).
[15] Y. Bramoullé, S. Currarini, M. O. Jackson, P. Pin, and B. W. Rogers, J. Econ. Theory **147**, 1754 (2012).
[16] J. Yun, P.-J. Kim, and H. Jeong, PLoS ONE **10**, e0117388 (2015).
[17] V. D. P. Servedio, G. Caldarelli, and P. Buttà, Phys. Rev. E **70**, 056126 (2004).
[18] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins, in *Proceedings of the 5th Annual International Conference on Computing and Combinatorics, COCOON'99* (Springer-Verlag, Berlin, 1999), pp. 1–17.
[19] A. Vazquez, Europhys. Lett. **54**, 430 (2001).
[20] P. L. Krapivsky and S. Redner, Phys. Rev. E **71**, 036118 (2005).
[21] F.-X. Ren, H.-W. Shen, and X.-Q. Cheng, Physica A **391**, 3533 (2012).
[22] M. O. Jackson and B. W. Rogers, Am. Econ. Rev. **97**, 890 (2007).
[23] S. R. Goldberg, H. Anthony, and T. S. Evans, Scientometrics **105**, 1577 (2015).
[24] Z.-X. Wu and P. Holme, Phys. Rev. E **80**, 037101 (2009).
[25] R. Itzhack, L. Muchnik, T. Erez, L. Tsaban, J. Goldenberg, S. Solomon, and Y. Louzoun, Physica A **389**, 5308 (2010).
[26] T. Martin, B. Ball, B. Karrer, and M. E. J. Newman, Phys. Rev. E **88**, 012814 (2013).
[27] D. M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles, Proc. Natl. Acad. Sci. USA **99**, 5207 (2002).
[28] W. Goffmann and V. A. Newill, Nature (London) **204**, 225 (1964).
[29] E. Bruckner, W. Ebeling, and A. Scharnhorst, Scientometrics **18**, 21 (1990).
[30] L. M. Bettencourt, A. Cintran-Arias, D. I. Kaiser, and C. Castillo-Chavez, Physica A **364**, 513 (2006).
[31] F. M. Bass, Manage. Sci. **50**, 1825 (2004).
[32] N. K. Vitanov and M. R. Ausloos, in *Models of Science Dynamics*, edited by P. A. Scharnhorst and K. Borner (Springer, Heidelberg, 2012), pp. 69–125.
[33] Z.-G. Shao, X.-W. Zou, Z.-J. Tan, and Z.-Z. Jin, J. Phys. A: Math. Gen. **39**, 2035 (2006).
[34] G. J. Peterson, S. Presse, and K. A. Dill, Proc. Natl. Acad. Sci. USA **107**, 16023 (2010).
[35] A. Vazquez, Phys. Rev. E **67**, 056104 (2003).
[36] T. S. Evans and J. P. Saramaki, Phys. Rev. E **72**, 026138 (2005).
[37] D. J. de Solla Price, Science **149**, 510 (1965).
[38] K. Borner, J. T. Maru, and R. L. Goldstone, Proc. Natl. Acad. Sci. USA **101**, 5266 (2004).
[39] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.95.012324 for details of our measurements.
[40] This definition of direct or indirect references differs from that of Ref. [34] and is more close to fresh or old papers of [49], random or ancestor search of [20], broadcasting or word-of-mouth search of [32], adding or walking of [19,35].
[41] N. L. Geller, J. S. de Cani, and R. E. Davies, J. Am. Soc. Inf. Sci. **32**, 1 (1981).
[42] E. R. Stinson and F. Lancaster, J. Inf. Sci. **13**, 65 (1987).
[43] H. Nakamoto, Informetrics **87/88**, 157 (1988).
[44] S. Redner, arXiv:physics/0407137.
[45] C. Roth, J. Wu, and S. Lozano, J. Inf. **6**, 111 (2012).
[46] W. Glanzel, Scientometrics **60**, 511 (2004).
[47] H. Bouabid and V. Larivière, Scientometrics **97**, 695 (2013).
[48] B. K. C. Parongama Sen, *Sociophysics: An Introduction* (Oxford University Press, New York, 2014).
[49] M. V. Simkin and V. P. Roychowdhury, J. Am. Soc. Inf. Sci. Technol. **58**, 1661 (2007).

[50] J. R. Clough, J. Gollings, T. V. Loach, and T. S. Evans, J. Complex Netw. **3**, 189 (2014).
[51] P. Chen, H. Xie, S. Maslov, and S. Redner, J. Inf. **1**, 8 (2007).
[52] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, Comput. Netw. **33**, 309 (2000).
[53] M. Golosovsky and S. Solomon, Phys. Rev. Lett. **109**, 098701 (2012).
[54] S. Zhou and R. J. Mondragon, Phys. Rev. E **70**, 066108 (2004).
[55] M. Rosvall, A. V. Esquivel, A. Lancichinetti, J. D. West, and R. Lambiotte, Nat. Commun. **5**, 4630 (2014).
[56] T. E. Harris, *The Theory of Branching Processes* (Springer-Verlag, Berlin, 1963).
[57] W. Ebeling, A. Engel, and V. G. Mazenko, Biosystems **19**, 213 (1986).
[58] R. Crane and D. Sornette, Proc. Natl. Acad. Sci. USA **105**, 15649 (2008).
[59] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, in *Proceedings of the 23rd International Conference on World Wide Web* (Association for Computing Machinery, New York, 2014).
[60] J. L. Iribarren and E. Moro, Phys. Rev. E **84**, 046116 (2011).
[61] M. Medo, Phys. Rev. E **89**, 032801 (2014).
[62] G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Muñoz, Phys. Rev. Lett. **89**, 258702 (2002).
[63] R. Cohen, S. Havlin, and D. ben-Avraham, Phys. Rev. Lett. **91**, 247901 (2003).
[64] M. J. Stringer, M. Sales-Pardo, and L. A. N. Amaral, PLoS ONE **3**, e1683 (2008).
[65] B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones, Science **342**, 468 (2013).
[66] D. E. Acuna, S. Allesina, and K. P. Kording, Nature (London) **489**, 201 (2012).
[67] A. Mazloumian, PLoS ONE **7**, e49246 (2012).
[68] O. Penner, R. K. Pan, A. M. Petersen, K. Kaski, and S. Fortunato, Sci. Rep. **3**, 3052 (2013).
[69] M. E. J. Newman, Europhys. Lett. **105**, 28002 (2014).
[70] I. V. Ponomarev, D. E. Williams, C. J. Hackett, J. D. Schnell, and L. L. Haak, Technol. Forecasting Soc. Change **81**, 49 (2014).
[71] Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini, Proc. Natl. Acad. Sci. USA **112**, 7426 (2015).
[72] T. Onaga and S. Shinomoto, Phys. Rev. E **89**, 042817 (2014).
[73] T. S. Kuhn, *The Structure of Scientific Revolutions* (University of Chicago Press, Chicago, 1970).
[74] A. Ishii, H. Arakaki, N. Matsuda, S. Umemura, T. Urushidani, N. Yamagata, and N. Yoshida, New J. Phys. **14**, 063018 (2012).
[75] J. Leskovec, L. A. Adamic, and B. A. Huberman, ACM Trans. Web **1**, 5 (2007).
[76] R. Pastor-Satorras, C. Castellano, P. V. Mieghem, and A. Vespignani, Rev. Mod. Phys. **87**, 925 (2015).
[77] D. Centola, V. M. Eguíluz, and M. W. Macy, Physica A **374**, 449 (2007).
[78] D. Centola, Science **329**, 1194 (2010).
[79] F. J. Pérez-Reche, J. J. Ludlam, S. N. Taraskin, and C. A. Gilligan, Phys. Rev. Lett. **106**, 218701 (2011).
[80] D. Broder-Rodgers, F. J. Pérez-Reche, and S. N. Taraskin, Phys. Rev. E **92**, 062814 (2015).
[81] J. J. Ludlam, G. J. Gibson, W. Otten, and C. A. Gilligan, J. R. Soc. Interface **9**, 949 (2011).
[82] G. Csárdi, K. J. Strandburg, L. Zalányi, J. Tobochnik, and P. Érdi, Physica A **374**, 783 (2007).
[83] S. Valverde, R. V. Solé, M. A. Bedau, and N. Packard, Phys. Rev. E **76**, 056118 (2007).
[84] P. Sheridan, Y. Yagahara, and H. Shimodaira, Physica A **391**, 5031 (2012).
[85] J. P. Bagrow and D. Brockmann, Phys. Rev. X **3**, 021016 (2013).