Growing complex network of citations of scientific papers- measurements and modeling- Supplementary Material

Michael Golosovsky and Sorin Solomon
The Racah Institute of Physics,
The Hebrew University of Jerusalem, 91904 Jerusalem, Israel

I. LOGICAL ORDER OF PRESENTATION

I: Introduction

II: Citation dynamics from the author's perspective

- **A:** Recursive search algorithm
- **B:** Model: age distribution of references
- **C:** Measurements: direct and indirect references
  1. Reduced age distribution of references
  2. Measurement of $R_{dir}(t), R_{indir}(t)$
  3. Dependence on the publication year

III: Reference-citation duality

- **A:** Duality
- **B:** Mean-field citation dynamics

IV: Citation dynamics from the perspective of cited paper

- **A:** Model: citation dynamics of individual papers
- **B:** Methodology: citation dynamics of the groups of similar papers
- **C:** Direct citations
- **D:** Indirect citations
- **E:** Second-generation citations and citing papers
- **F:** Probability of indirect citation

V: Stochastic model of citation dynamics

- **A:** Model formulation
- **B:** Model validation
  1. Methodology
  2. Citation distributions
  3. Citation trajectories
  4. Short summary

VI: Discussion

- **A:** Comparison to previous studies
- **B:** Nonlinearity and "power-law" degree distributions
- **C:** Preferential attachment
- **D:** Prediction of citation trajectories of papers
- **E:** Comparison to other research fields
- **F:** Extension to other networks

VII: Summary

Appendix A: A hand-waving explanation of the nonlinear probability of indirect citation

Appendix B: Our results in the context of network science

The paper is organized as follows. Section I, Introduction, stresses the need for the comprehensive model of citation dynamics that is validated in several dimensions by dedicated measurements. Section II focuses on references. We propose there a plausible scenario that the authors follow when they compose the reference lists of their papers. Basing on 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 establishes the reference-citation duality. Basing on this duality and the model for the age distribution of references of Section II A-B we develop a mean-field model of citation dynamics. This macroscopic approach captures the mean citation dynamics of a single research field. In Section IV we "individualize" the mean-field model of Section III B to capture 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. The model contains empirical functions which we find in dedicated measurements (Section IV C,D). Due to reference-citation duality (Section III A) we expect a certain relation between these empirical functions and those found in our studies of references (Section II C2). We verify these relations and find that some of them do not hold, namely, the probability of indirect citations is not the same for all papers. This prompts us to consider in more detail the second-generation citations (Section IV E) and to study in detail the probability of indirect citations (Section IV F). Thus we are fully equipped to develop the model of citation dynamics of individual papers. Section V deals with stochastic model of citation dynamics of individual papers. This represents a truly microscopic approach and it is the main message of the paper. Section V B focuses on model validation in several dimensions, the Supplementary Material contains much more material related to model validation. Section VI briefly discusses implications of our study for network science.

II. DATASETS

- *21 PRB papers published in 2014.* This set was used to found the age composition of references: $R(t), R_{dir}(t)$ and $R_{indir}(t)$.

  We chose a small but representative set of papers in one field and analyzed their reference lists. Namely, we made an unbiased choice of 21 research papers published in the Physical Review B in 2014 (a mix of theoretical and experimental papers that cover several different research subfields) and analyzed their first-generation and second-generation references using Scopus database (approximately 600 first-generation references and 18000 second-generation references). To distinguish between the direct and direct references we made the following. For each parent paper we compared the lists of its first- and second-generation references and identified the papers that appear in both lists. These are indirect references. The papers that appear only in the first but not in the second list are direct references. We arranged the unified first-generation reference list of these 21 parent papers in chronological order, counted the numbers of direct and indirect references published in each year, and divided them by the total number of references. This yields $R(t), R_{dir}(t)$ and $R_{indir}(t)$.

- *PRB papers published in the July issues of 1998, 2004, 2014.* This set was used to find the reduced reference age, $r(t)$.

  We choose one issue of the Physical Review B and considered all original research papers published there. We identified the references of these papers using Scopus database, arranged them in the chronological order and found the number of papers published in each year. We divided this number by the total number of references and found $r(t)$. This was done for years: 1998, 2004, and 2014. The sample size is 4373 papers in 1998, 4964 in 2004, and 4813 in 2014 (only articles and letters). For all three years we focused on July issue, since it is published in the middle of the year.

- *2078 Physical Review B papers published in 1984.* These were used to compare citation and reference list length distributions.

  The set contains all original research PRB papers published in 1984. Citations were counted in 2014.

- *37 PRB papers published in 1984.* These were used for measuring direct and indirect citations, $K_{dir}(t)$ and $K_{indir}(t)$.

  We focused on one research field- Physics, one research journal- Physical Review B, and one publication year - 1984. We performed our analysis manually and selected small but representative groups of original research papers that garnered the same number of citations K_∞ by the end of 2013 (14 papers with 10 citations, 10 papers with 20 citations, 10 papers with 30 citations, and 3 papers with 100 citations, the mix of theoretical and experimental papers). We measured citation dynamics of these 37 papers using the Thomson-Reuters Web of Science database. For each parent paper we considered the list of first- and the second-generation citing papers (overviews excluded, self-citations included). We identified indirect citations as those that appear in both lists. The direct citations are those that appear only in the first list.

- *108 PRB papers published in 1984.* These were used to measure statistics of the second-generation citations and citing papers.

  We considered 108 Physics papers published in the Physical Review B in 1984 and arranged them into several groups, each of which consisting of papers that garnered approximately the same number of citations K_i by the end of 2013 i.e., 30 years after publication. (22 papers with 10 citations, 19 papers with 20 citations, 15 papers with 30 citations, 12 papers with 48-52 citations, 21 papers with 95-105 citations, 11 papers with 210-240 citations, and 8 papers with 470-560 citations). For each source paper i we counted the total number of its second-generation citations and citing papers that were published by the end of 2013, and divided these counts by K_i. This yields M_i^{nn} and N_i^{nn}, correspondingly. Then we calculated $M^{nn} = \overline{M_i^{nn}}$ and $N^{nn} = \overline{N_i^{nn}}$, the average over each group of papers with the same K_i.

- *3 PRB papers published in 1984.* These were used to measure probability of indirect citation.

  We chose three representative Physical Review B papers that were published in 1984 and gained 100 citations by the end of 2013. [F. T. Gittes and M. Schick, *Complete and incomplete wetting by adsorbed solids*, PRB 30, pp. 209-214 (1984); F. J. Himpsel, P. M. Marcus, R. Tromp, Inder P. Batra, M. R. Cook , F. Jona, and H. Liu, *Structure analysis of Si(111)2x1 with low-energy electron diffraction*, PRB 30, pp. 2257-2259 (1984); Wan Y. Shih and D. Stroud, *Superconducting arrays in a magnetic field: Effects of lattice structure and a possible double transition*, PRB 30, pp. 6774–6777 (1984)].

  We studied two generations of their citing papers while limiting ourselves only to descendants of the direct citations and disregarding indirect citations bringing another indirect citation. For each parent paper we pinpointed direct citations (first generation) and the papers that cite them (second generation). These data were analyzed using the protocol described in the paper. (One of these three papers was cited by a very important overview that in its turn was cited more than 3000 times. We excluded this overview from our analysis of the second-generation citations).

- *40,195 Physics papers published in 1984.* These were used to compare measured and simulated citation distributions.

  We chose 83 leading Physics journals published in 1984 and measured citation trajectories of the original research papers published there (overviews excluded, self-citations included) using Thomson-Reuters Web of Science database (We already used this set of papers in our previous studies). To this end we downloaded citation histories of these papers and analyzed them using MATLAB software. The downloading was performed manually, by the groups of 500 papers using Citation Report option. Then we analyzed citation distributions, citation trajectories of different papers, Pearson autocorrelation coefficient for annual citations, citation lifetime, etc.

- *48,168 papers Physics papers published in 1984.* These were used to measure $M(t)$, the mean citation rate of Physics papers.

  We chose all original research Physics papers published in 1984 (overviews excluded, self-citations included) using Thomson-Reuters Web of Science database. We didn't download all citations of these papers but measured their mean annual citation rate up to 2013 using Citation Report option of the database.

III. NUMBER OF PUBLICATIONS

To find how the number of publications depends on time we measured $N_0(t_0)$, the total number of Physics papers published annually during the period 1980-2013. We used the Thomson-Reuters Web of Science database. Figure 1 shows exponential growth $N_0(t_0) \propto e^{\alpha t_0}$ with the exponent $\alpha = 0.046$ (2% annual growth) consistent with previous estimates of the growth of Physics publications in the period 1980-2010 [5, 6].

To find how the number of references per paper depends on time we measured $R_0(t_0)$, the average length of the reference list of the Physical Review B papers published in 1996-2013. For this measurement we used the Scopus database. Figure 1 shows a very weak dependence on the publication year which we approximated by an exponential, $R_0 \propto e^{\beta t}$. Ref. [7] performed similar measurements for the Physical Review papers (only PR to PR references were

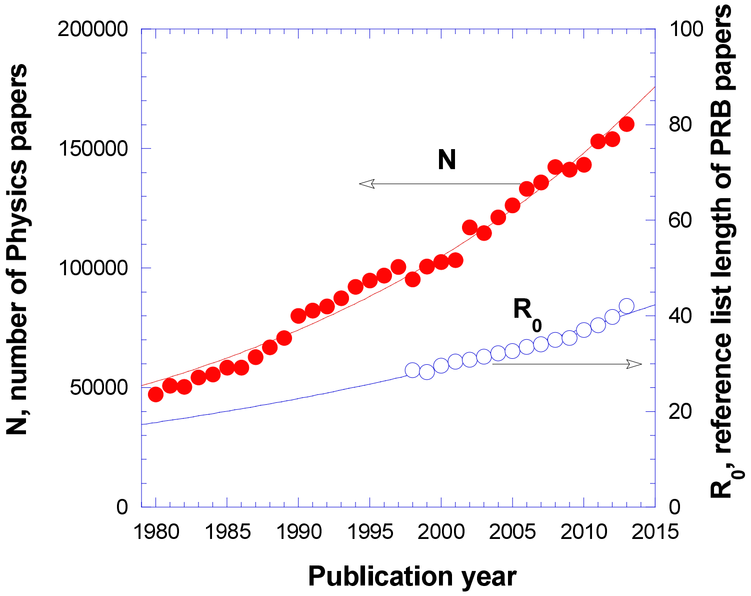

FIG. 1. Time dependence of the number of Physics papers published in 1980-2013 (full red circles). The red continuous line shows exponential approximation $N_0(t_0) \propto e^{\alpha t_0}$ where $\alpha = 0.046$ yr^{-1}. The empty circles show time dependence of the average length of the reference list of a Physical Review paper. The blue continuous line shows exponential approximation, $R_0 \propto e^{\beta t_0}$ where $\beta = 0.02$ yr^{-1}.

considered) and approximated their results by the logarithmic dependence. Ref. [8] also performed such measurements and claimed that there is a very slow growth of the reference list length before 2000 and subsequent acceleration following the advent of open access and electronic format journals that have no page limit. Although $R_0(t_0)$ was approximated by very different functional dependences, our measurements of $R_0(t_0)$ and the measurements of other authors [6, 7] for the Physical Review papers are very close.

IV. AGE COMPOSITION OF THE REFERENCE LIST, $R(t)$

To measure $R(t)$ we took ensemble of papers published in the July issue of the Physical Review B and downloaded their lists of references using SCOPUS database. Then we arranged all these lists together, measured the number of references published in every year, and divided this number by the total number of the papers. This yields $R(t)$. This procedure was performed for: all 444 papers published in Physical Review B 90, issues 1-4 (July 2014); all 379 papers published in Physical Review B 70, issues 1-4 (July 2004); and all 323 papers published in Physical Review B 58, issues 1-4 (July 1998).

The download was not performed on the paper-by-paper basis since it was too time-consuming. Instead, we decided to download batches of papers although we were fully aware that there is a potential pitfall here. Indeed, all papers in one batch pertain to the same field - Physics, hence there is a chance that their references overlap. If we download the reference lists of several papers together and there are overlapping references, the SCOPUS counts only one of them. Thus, to minimize this overlap we have to use small batches. To study the probability of such overlap we performed dedicated measurements and found that for the batches of 20 papers the overlap is small, only 1.9%; while for larger batch (444 papers) the overlap is quite significant- 9.2%. In fact, this overlap may be not an issue since overlapping references may be distributed uniformly over years. However, we didn't take chance and in our downloading procedure we used small batches of 20 papers (a minimal page in SCOPUS).

V. WHAT IS THE FRACTION OF INDIRECT REFERENCES IN THE REFERENCE LISTS OF PAPERS?

Our study of the reference lists of 21 Physical Review B papers revealed that the reference lists of these papers include 65% indirect references and 35% direct references. Previous estimates yielded: Refs. [1, 2] report, correspondingly, 67-78% and 80% indirect references in the reference lists of high-energy Physics preprints; Ref. [3] reports 56.4% indirect references in the American Physical Society publications (only APS to APS references were counted); Ref. [4] found 40-50 % indirect references in the Physical Review publications (only PR to PR references were counted).

There is some difference between these measurements and ours. We attribute this difference to the fact that almost all above studies measured citations, not references, the tacit conclusion being that the average fraction of indirect references for some set of papers is the same as the average fraction of indirect citations. Our studies show that this is not true and the fraction of indirect citations of a paper depends on the total number of citations it garnered, in such a way that this fraction increases from 30% for low-cited papers to 80 % for highly-cited papers. Hence, the estimate of the fraction of indirect references through the measurement of the fraction of indirect citations for the set of papers that contain both low- and highly-cited papers is problematic.

VI. THE FITTING PROCEDURE TO FIND MODEL PARAMETERS

We measured $R(t)$ and $R_{indir}(t)$, the numbers of all references and indirect references in the reference list of papers. The model yields the following relation between them

$$R_{indir}(t) = \sum_{\tau=0}^{t} R(t-\tau)P(\tau)e^{-\beta\tau}R(\tau). \tag{1}$$

where $P(\tau)$ is the probability of indirect citation and $e^{-\beta\tau}$ characterizes the time dependence of the reference list length ($\beta = 0.02$ yr^{-1}). Basing on Eq. 1 we found P_0 and γ by the fitting procedure. Namely, we assumed that $P(\tau) = P_0 e^{-\gamma\tau}$ and we looked for the values of P_0 and γ that produce the best fit. We were able to achieve good fits for several combinations of P_0 and γ. By trial and error we found that if we cast the above expression as follows: $P(\tau) = \Pi\gamma e^{-\gamma\tau}$ where $\Pi = P_0/\gamma$, and used Π and γ as fitting parameters instead of P_0 and γ, then the fit is sensitive to Π and not very sensitive to γ. Indeed, for the same $\Pi = 6.3$ we found reasonable fits for $\gamma = 1 - 1.8$ yr^{-1}. Thus, basing on Eq. 1 we can find Π but can't reliably determine γ. We looked closely for another measure of citation dynamics that is more sensitive to γ and found that it is the Pearson autocorrelation coefficient $c_{t,t-1}$ for additional citations (Eq. 6). For given Π we varied γ to achieve the best correspondence between the measured and numerically simulated $c_{t,t-1}$ as it is shown in Fig. 5. The best fit was achieved for $\gamma + \beta = 1.2$ yr^{-1}.

VII. CITATION DISTRIBUTION OF THE PHYSICAL REVIEW B PAPERS REPRESENTS THE WHOLE PHYSICS FIELD

A considerable part of our measurements has been performed using the papers published in the Physical Review B. To what extent these measurements are generic, namely, do citation patterns of the PRB papers represent Physics? To this end we compared cumulative citation distributions of the PRB papers and of all Physics papers published in 1980-1989. Figure 2 shows that although a PRB paper garners ~ 40 % more citations than an average Physics paper, citation distributions for the PRB papers and the whole Physics field are very similar.

Dynamics of direct and indirect citations for the PRB papers are also generic. Figure 3 shows citation dynamics of a PRB paper. The direct citations shoot up soon after publication and then their growth slows down while indirect citations appear after 1-2 year delay. Surprisingly, direct citations do not come to saturation even after 28 years.

VIII. DETAILS OF NUMERICAL SIMULATION

The following Equation summarizes our model

$$\lambda_i(t) = \eta_i m_{dir}(t) + \sum_{\tau=0}^{t} N^{nn}(t-\tau)\tilde{P}_0(K_i)e^{-(\gamma+\beta)(t-\tau)}k_i(\tau). \tag{2}$$

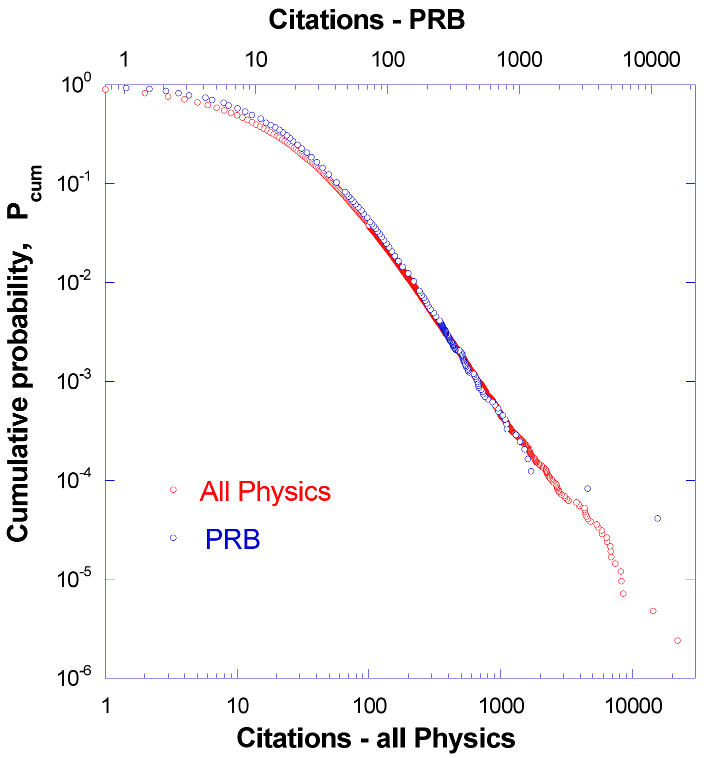

FIG. 2. Cumulative citation distribution for 418,438 Physics papers published in 1980-1989. Citations were counted in July 2008. Blue points show corresponding distribution for all Physical Review B papers published in 1984. Although these distributions differ in scale, their shapes are almost identical.

Here, $\lambda_i(t)$ is the latent citation rate of a paper i at year t after publication, η_i is paper's fitness, an empirical parameter, unique for each paper; $m_{dir}(t)$ 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, $Poiss(k_i) = \frac{\lambda_i^{k_i}}{k_i!}e^{-\lambda_i}$. The exponents γ and β, the functions $m_{dir}(t)$, $N^{nn}(t-\tau)$, and $\tilde{P}_0(K)$ are the same for all papers in one field published in one year.

In our numerical simulations of citation dynamics of 40,195 Physics papers published in 1984 we used the following parameters. $\gamma+\beta = 1.2$ yr^{-1}, as found in our measurements of indirect references and citations; $m_{dir}(t)$ from the Fig. 10 (see main text). We assumed that $N^{nn}(t)$ dependence mimics $M(t)$ dependence, namely $N^{nn}(t) = \frac{M(t)}{\bar{s}}$ where $\bar{s} = 1.2$ is the average over all Physics papers published in 1984 and $M(t)$ is shown in Fig. 6 (main text). With respect to $\tilde{P}_0(K)$, it is given by the following expression: $\tilde{P}_0(K) = Pf(K)$ where $P = 0.34$ and $f(K) = 1+0.82\log K$. Here, $P = 0.34$ characterizes the probability of indirect citations for low-cited papers, and $f(K)$ stays for logarithmic correction which is most important for highly-cited papers.

We can assmeble all time-dependent functions in the kernel together, in such a way that Eq. 2 reduces to

$$\lambda_i(t) = \eta_i m_{dir}(t) + \sum_{\tau=1}^{t} f(K_i)F(t-\tau)k_i(\tau) \quad (3)$$

where $F(t) = PN^{nn}(t)e^{-(\gamma+\beta)t}$. For the Physics papers published in 1984 we find $F(1) = 0.089, F(2) = 0.138, F(3) = 0.046, F(4) = 0.012, F(5) = 0.0035...$ Thus, at first F grows with time as the paper receives more recognition (there is approximately one year delay between the publication of the paper and its first citation) and then decays exponentially. This obsolescence is strong, hence the memory of the citation process is restricted to a few years.

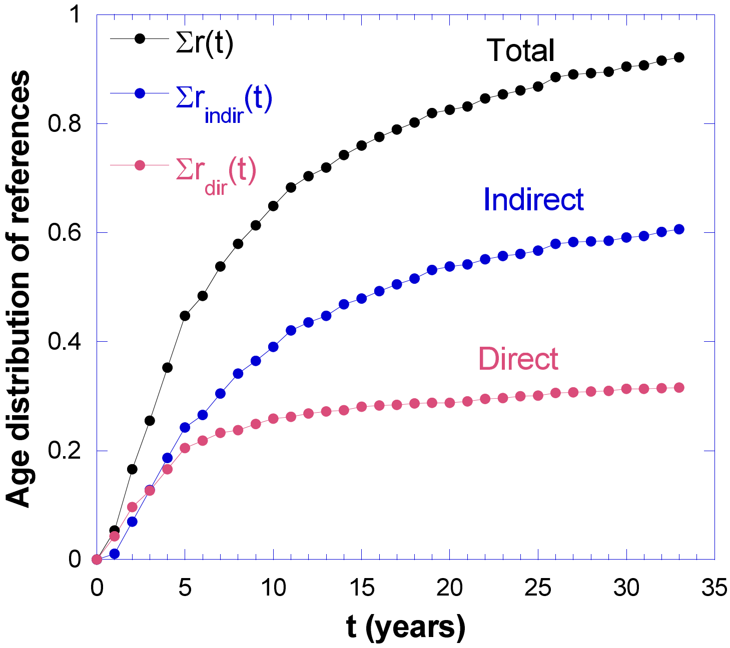

FIG. 3. Cumulative number of citations of a representative Physics paper (F. Himpsel et al., Phys. Rev. B **30**, 2257 (1984)). The direct citations shoot up immediately after publication of the parent paper and then their rate slowly decays. The indirect citations shoot up with 1-2 year delay, their rate achieves its maximum in another couple of years and then decays.

In our numerical simulations we used Eq. 3. For approximate calculations one can also use a simplified numerical scheme according to which Eq. 3 is considered as an autoregressive model. We looked for the model of minimum order that can faithfully represent our measurements and found that the first-order model is unsatisfactory while the second-order autoregressive model

$$\lambda_i(t) = \eta_i m_{dir}(t) + [1 + 0.82 \log K_i(t)][0.09 k_i(t) + 0.19 k_i(t-1)] \tag{4}$$

is more or less satisfactory. This means that to approximately predict the number of citations that a paper garners in the year $t+1$ after publication, in most cases it is enough to know its citation history during previous couple of years: t and $t-1$. Thus our results validate the widespread use of the two-year impact factor.

IX. STOCHASTIC MODEL VERIFICATION

Although the agreement between the measured and simulated citation distributions is impressive, this fact alone is not sufficient to validate our model. While early models of complex networks growth were validated by comparing measured and simulated aggregate characteristics, such as degree distribution, our model is the next-generation one, it is much more detailed and the comparison to real data is more demanding. To the best of our knowledge, the methodology of comparing stochastic model/simulation to stochastic data is not well-established. Following Ref. [9] we believe that the proper validation of stochastic models shall include multidimensional analysis where several predictions of the model are compared to measurements. In the text of our paper we demonstrate that our model reproduces cumulative citation distributions fairly well. In what follows we verify our model in several other dimensions.

A. Stochastic component of the citation dynamics

Microscopic citation dynamics is usually considered in relation to the preferential attachment mechanism which in its most general form is described by the following equation

$$\lambda_i(t) \propto [K_i(t) + K_0]^\delta, \tag{5}$$

where $K_i(t)$ is the number of citations of the source paper i garnered by year t, K_0 is the so-called initial attractivity, and δ is the growth exponent. Along this line of thinking, we measured citation dynamics of Physics papers using the set of dependent and independent variables suggested Eq. 5. In particular, for each t we sorted the papers into bins containing those with the same $K(t)$, the number of citations garnered by the time t. We considered distribution of additional citations k garnered by the papers in each bin in the year $t+1$ and calculated the mean and the variance of this distribution. This was done both for measured and simulated data.

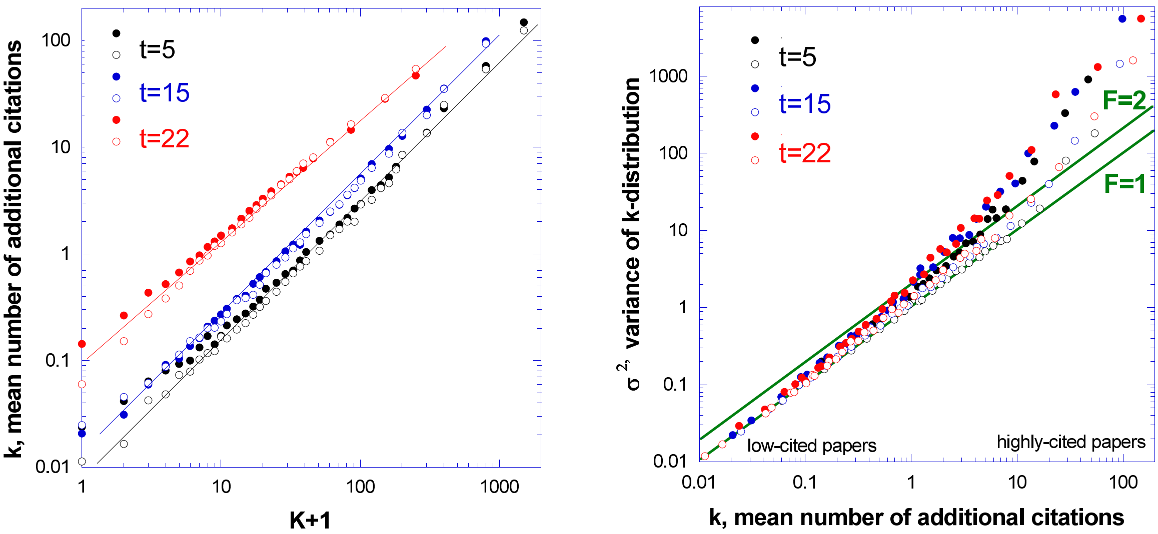

FIG. 4. (a) Mean number of additional citations $\overline{k_i(t)}$, in dependence of the number of previous citations $K(t)$. t is the number of years after publication. The averaging is performed over all papers published in one year that got the same number of citations K during t years after publication. The straight lines are fits to Eq. 5 with $K_0 = 1$. (b) Variance of the additional citations, $\sigma^2(t) = \overline{\left(k_i(t) - \overline{k_i(t)}\right)}$, versus mean, $\overline{k_i(t)}$. The filled circles show measured values, the empty circles show results of numerical simulation, the straight lines indicate constant variance-to-mean ratio (Fano number) where $F = 1$ corresponds to the Poisson distribution. The data are for 40,195 Physics papers published in 1984.

Figure 4a shows that the mean number of additional citations k for the measured and simulated data are very close: both follow Eq. 5 with $\delta \sim 1.25$ and $K_0 = 1$. Figure 4b plots the variance versus mean for these distributions. The rationale for such plot is the fact that for the Poisson distribution, the variance-to-mean ratio (Fano number) is $F = 1$, while for many other distributions $F > 1$. Hence, any deviation from the Poisson distribution can be easily noticed.

For small $\overline{k_i(t)}$ the measured and simulated data are close to one another and to the $F = 1$ line. This demonstrates a good agreement between the measurements and the model. It also means that the stochastic component of citation dynamics is Poissonian, namely random. For large $\overline{k_i(t)}$ the measured data deviate upwards from the $F = 1$ line. This means that the variability of citation dynamics of these highly-cited papers arises more from the differences in their citation history than from chance. Although the simulated data for highly-cited papers also deviate upwards from the $F = 1$ dependence, this deviation is smaller that that for measured data. Hence, our model captures well the mean citation dynamics of all papers, correctly predicts the variability of citation dynamics of the low- and moderately-cited papers, and underestimates it for highly-cited papers.

B. Autocorrelation

Another point of comparison is the autocorrelation of additional citations acquired by a paper in subsequent years. We characterize it by the Pearson autocorrelation coefficient, $c_{t,t-1}$. We calculated it for annual citations measured for the sets of papers which have the same number of previous citations $K(t)$. Specifically, we determined the number of citations garnered by each paper during two subsequent years, t and $t-1$: $k_i(t)$ and $k_i(t-1)$, correspondingly. Then we calculated the Pearson autocorrelation coefficient

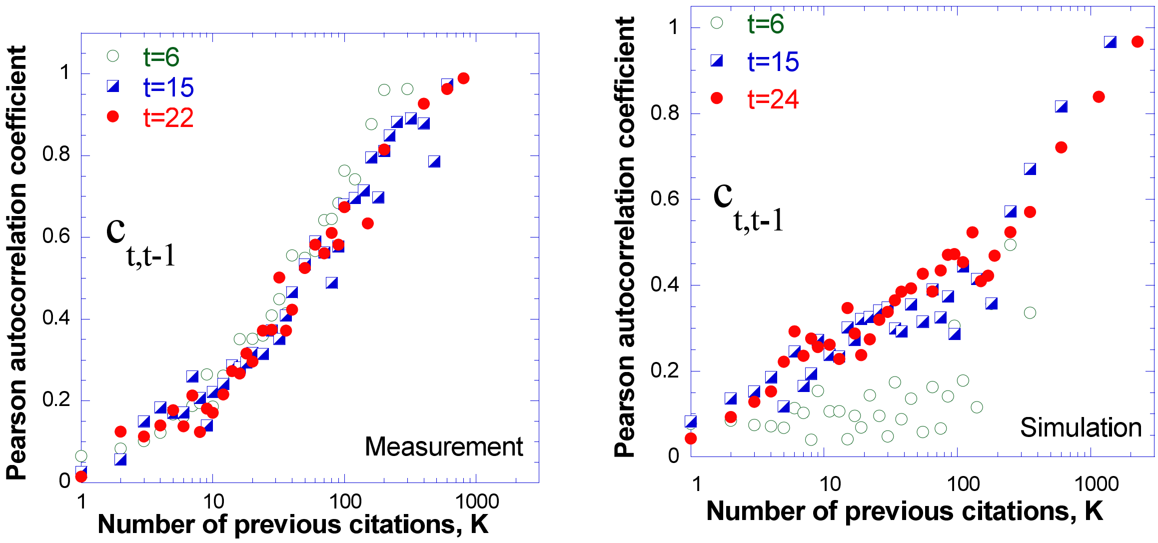

FIG. 5. The Pearson autocorrelation coefficient, $c_{t,t-1}$, for additional citations in subsequent years, $k_i(t)$ and $k_i(t-1)$. Each point corresponds to the set of papers with the same number of previous citations K garnered during t years after publication. (a) Measurements. (b) Numerical simulation. The simulation agrees with the measurements for $t > 10$ yr and underestimates $c_{t,t-1}$ for $t < 10$.

$$c_{t,t-1} = \frac{\overline{\left(k_i(t) - \overline{k_i(t)}\right)\left(k_i(t-1) - \overline{k_i(t-1)}\right)}}{\sigma_t \sigma_{t-1}} \qquad (6)$$

where, σ_t, σ_{t-1} are, correspondingly, the standard deviations of the $k_i(t)$ and $k_i(t-1)$ distributions, and the averaging is performed over all papers in the set.

Figure 5 shows our results. We do not know why $c_{t,t-1}(K)$ dependences for different years collapse onto a single curve. A more important fact is that $c_{t,t-1}$ grows with K. This is a direct consequence of the nonlinear $\tilde{P}_0(K)$ dependence discussed in the main text. Our model reproduces this growth fairly well for $t > 10$. For $t < 10$ the simulated values of $c_{t,t-1}(K)$ are lower than measured ones. This discrepancy can be lifted by assuming that the obsolescence exponent γ weakly depends on K. We reserve this topic for future studies.

What is the meaning of $c_{t,t-1}$? Low $c_{t,t-1}$ indicates that the stochastic component of citation dynamics is random, high $c_{t,t-1}$ indicates that it is determined by previous history. Consequently, small $c_{t,t-1}$ is associated with jerky, and $c_{t,t-1} \sim 1$ is associated with smooth citation trajectories. In the paper we compared the measured and numerically simulated citation trajectories of the Physics papers that were published in 1984. For moderately-cited papers the measured and simulated trajectories look similar- the fluctuations are of the same size and the spread in trajectories is the same. Trajectories are jerky, consistent with low $c_{t,t-1}$. For highly-cited papers both measured and simulated trajectories are smooth, that is consistent with high $c_{t,t-1}$. However, the spread of the measured trajectories exceeds the spread of the simulated ones.

C. Uncited papers

Figure 6 shows that our model correctly predicts the number of uncited papers at every year after publication. We found that only a small fraction -7.5% of the Physics papers published in 1984 remained uncited after 25 years. The good correspondence between the measured number of uncited papers and the model prediction indicates that uncited papers are a natural outcome of citation process (provided it follows the Poisson statistics)[10], they are not unread and contribute to scientific progress being an integral part of scientific enterprise [11].

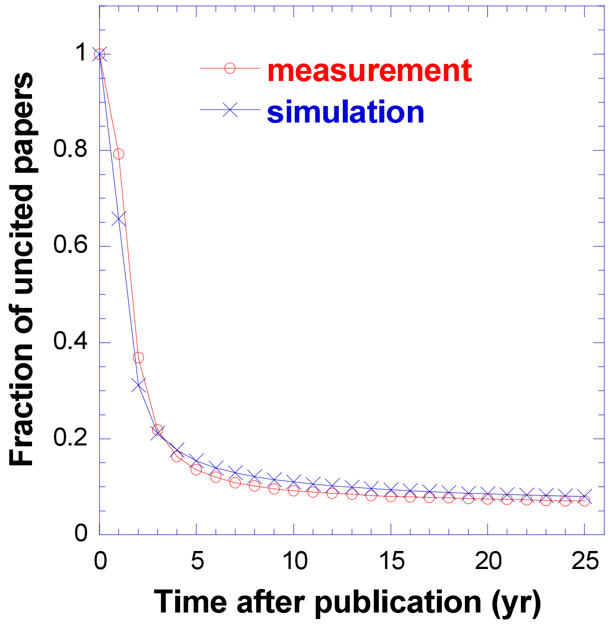

FIG. 6. Uncited papers. Time dependence of the fraction of the Physics papers that remained uncited 25 years after publication. The data is for the set of 40,195 Physics papers published in 1984. Note a good agreement between the measurements and simulations.

[1] J. R. Clough, J. Gollings, T. V. Loach, and T. S. Evans, Journal of Complex Networks **3**, 189 (2014).
[2] S. R. Goldberg, H. Anthony, and T. S. Evans, Scientometrics **105**, 1577 (2015).
[3] Y. Bramoullé, S. Currarini, M. O. Jackson, P. Pin, and B. W. Rogers, Journal of Economic Theory **147**, 1754 (2012).
[4] P. Chen, H. Xie, S. Maslov, and S. Redner, Journal of Informetrics **1**, 8 (2007).
[5] P. O. Larsen and M. von Ins, Scientometrics **84**, 575 (2010).
[6] T. Martin, B. Ball, B. Karrer, and M. E. J. Newman, Phys. Rev. E **88**, 012814 (2013).
[7] P. L. Krapivsky and S. Redner, Phys. Rev. E **71**, 036118 (2005).
[8] T. S. Evans, N. Hopkins, and B. S. Kaube, Scientometrics **93**, 473 (2012).
[9] M. Medo, Physical Review E **89**, 032801 (2014).
[10] Q. L. Burrell, Journal of Informetrics **7**, 676 (2013).
[11] P. O. Seglen, Journal of the American Society for Information Science **43**, 628 (1992).