Distribution for the number of coauthors

Jiann-wien Hsu[1] and Ding-wei Huang[2]
[1]*General Education Center, National Tainan Institute of Nursing, Tainan, Taiwan*
[2]*Department of Physics, Chung Yuan Christian University, Chung-li, Taiwan*
(Received 20 January 2009; revised manuscript received 10 August 2009; published 4 November 2009)

We study the coauthorship distribution by analyzing the number of coauthors on each paper published in *Physical Review Letters* and *Physical Review* for the last decade. We propose that the structure of the distribution can be understood as the result of a two-parameter Poisson process. We develop a dynamic model of dual mechanisms to simulate the personal and group collaborations. In this model, the single-author papers are portrayed as a leftover from the collaboration process. We also comment on the huge collaborations involving hundreds of coauthors.

DOI: 10.1103/PhysRevE.80.057101 PACS number(s): 89.65.−s, 01.30.−y, 89.75.Hc

I. INTRODUCTION

During the last century, science has evolved from individual research to teamwork. The collaboration among researchers has become an interesting topic to be explored with various approaches, such as the shift of social structure [1], the development of international relations [2], ethics in science [3], or a fair share of credit [4]. Within statistical physics, scientific collaboration is a typical example of social networks [5,6]. The interconnection of scientific community has been actively investigated [7–9], and the coauthorship has been used to define the connection between researchers [10]. The underlying network structure has been shown to have the properties of being small-world [11,12] and scale-free [13]. Most previous works placed a focus on the coauthorship networks. In this work, we report a study to explore the number of coauthors on a published paper, which would reveal a different aspect of the collaboration structure.

Figure 1 illustrates the typical distribution of coauthorship for the prestigious research journal, *Physical Review Letters*. Because the yearly distributions remain basically unchanged for the last decade, it is meaningful to accumulate all the data to achieve better statistics, which is shown by the steplike histogram in Fig. 1. We find that the distribution is dominated by the collaborations of two and three coauthors. As the number of coauthors increases, the structure of the distribution decreases monotonically. A two-slope structure is discerned: (1) a quick exponential decay appearing when the coauthors are fewer than 15 and (2) a large tail consisting of collaborations of over 20 coauthors. In this Brief Report, we focus on the distribution with the number of coauthors fewer than 100. A brief comment on the huge collaboration involving hundreds of researchers is presented in the discussions.

One of the basic criteria to warrant publication is that the paper contains sufficient new research. In such a sense, overlapping with an already published work might be the primary setback to a research paper submitted for publication. Substantially, many physical processes also involve independent events at a fundamental level, where the Poisson process provides a basic framework. Typical examples are the radioactive decay of atoms, incoming telephone calls at a switchboard, and passengers arriving at a bus stop, etc. We propose that the publication of research papers can also be taken as a Poisson process to understand the structure of the coauthorship distribution. Each researcher should be independent when in pursuit of a research topic and a publication schedule. When some of the researchers happen to share a common interest in a topic, they might wish to collaborate and establish the coauthorship in a published paper. We find that with this spontaneity of scholarly research, the collaboration among researchers can be described analytically as a Poisson process. The details are analyzed in Sec. II. In Sec. III, we further propose a cellular automaton model to numerically describe the dynamics of collaboration. Two different modes of collaboration are distinguished, which result in the two-slope structure in the coauthorship distribution. Our proposed model also addresses the depletion of single-author papers, leading to the dominance of two- and three-author papers in the distribution.

II. POISSON PROCESS

We propose that the Poisson process can be adapted to describe a process where multiple researchers collaborate to

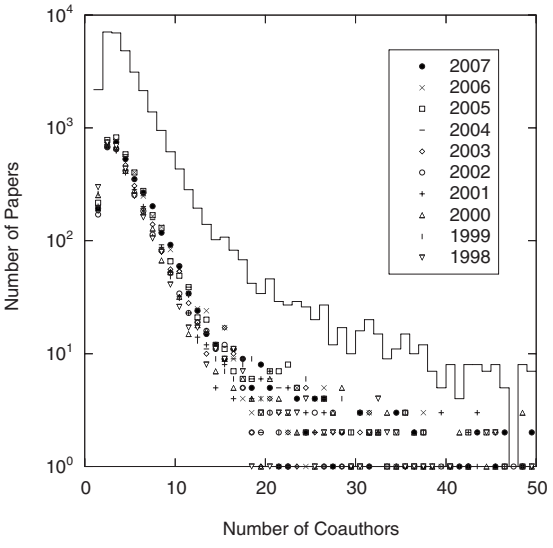

FIG. 1. Distributions of coauthorship in *Physical Review Letters*. The results in different years are shown by different symbols. The solid line shows the histogram for accumulated results in ten years.

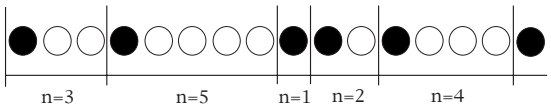

FIG. 2. Configuration of a virtual series. Each dot represents one author. Each paper starts with a first author, who is specified by a black dot. Nonfirst authors are represented by white dots. The configuration shows a three-author paper, followed by a five-author paper, a single-author paper, a two-author paper, and then a four-author paper, etc.

coauthor a published paper. Consider a principal investigator who proposes a research topic and naturally becomes the first author of the final publication. The principal investigator's proposal may attract other researchers to join the project, which eventually introduce the many coauthors of the published paper. Since every researcher has different reasons for choosing a topic or joining a project, it is plausible to consider these processes as being independent of each other.

On the other hand, at the journal editor's desk, research papers arrive continuously. After reviewed, some of the submitted papers are accepted for publication. In the following, we consider only the papers being accepted. The authors' name list in the contents of a published volume is taken as a virtual time series to reflect the above-mentioned collaborations as shown in Fig. 2. In this virtual process, each publication consists of a first author (black dot) followed by a few nonfirst authors (white dots). As expected, the appearance of the black dots is a simple stochastic process. It can be also expected that all the authors in the ensemble have the potential to be first authors. With a stochastic probability λ, an author will submit a new proposal; with the probability $(1-\lambda)$, the author may not submit the new proposal but instead join the existing one as a coauthor. The virtual series shown in Fig. 2 becomes a homogeneous Poisson process with the rate parameter λ. With a continuous approximation, the normalized distribution of coauthorship can then be written as

$$P(n) = (e^{\lambda} - 1)e^{-\lambda n}, \tag{1}$$

where n denotes the number of coauthors listed on a paper. This expression is valid for a small λ. The average number of coauthors becomes

$$\langle n \rangle = \frac{e^{\lambda}}{e^{\lambda} - 1} \sim \frac{1}{\lambda}. \tag{2}$$

The distribution of the number of coauthors in *Physical Review Letters* can be fairly reproduced by an incoherent sum of two Poisson processes as shown in Fig. 3. For $n < 15$, the data can be described with a large stochastic probability $\lambda_1 = 0.4$. For $n > 20$, the data can be fitted by prescribing a much smaller probability at $\lambda_2 = 0.07$. These observations suggest two distinct mechanisms behind the distribution. For the small-group collaboration involving fewer than 15 researchers, the steep decent of the distribution implies that each researcher has a strong wish to propose a new project. For the large-group collaboration involving more than 20 researchers, the gentle decent of the distribution can be related to a tendency among researchers to join an established

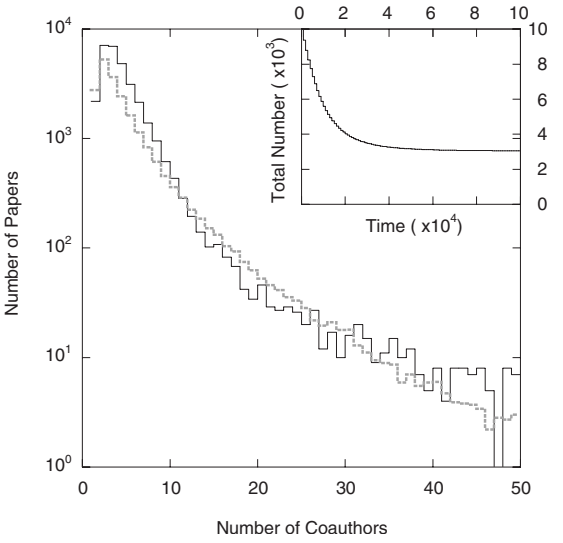

FIG. 3. Distribution of coauthorship in analytical Poisson process and numerical cellular automaton model. The two solid lines show the exponential decay with $e^{-0.4n}$ and $e^{-0.07n}$, where n denotes the number of coauthors. Their incoherent sum is shown by the gray-dashed line. The gray-dotted histogram shows the numerical results with $(\epsilon_1, \epsilon_2) = (0.4, 0.07)$. The data from *Physical Review Letters* are shown by the solid histogram. The inset shows the time evolution of total number of papers in the cellular automaton model.

project. With this simple description, the general shape of the coauthorship distribution can be understood, except for the amount of single-author papers. We find that the Poisson process leads to a false expectation that single-author papers should dominate the distribution.

III. CELLULAR AUTOMATA

In this section, we propose a simple dynamic model to study how researchers collaborate. We do not intend to have a detailed model examining the complex evolution of scientific collaboration. We aim to propose a basic model to present both the exponential decay in last section and the depletion of single-author papers. Consider 10^4 researchers located on a two-dimensional 100×100 regular lattice. Each lattice site represents an independent researcher. This ensemble of 10^4 researchers has roughly the same amount as the total number of authors appeared in the *Physical Review Letters* for each year [14]. A real number $a_i \in (0,1)$ is assigned to each lattice site representing the research topic chosen by the researcher on that site, where the index i runs through each and every site. The variable a_i can be taken as a continuum version of the Physics and Astronomy Classification Scheme (PACS) numbers, which provides a conventional scheme to specify a research topic.

Theoretically, the scope of *Physical Review Letters* covers all fields of physics. In addition, all the coauthors can be expected to contribute significantly to the published papers. It is then reasonable to assume that each coauthor has the potential to carry out independent research and to publish a single-author paper. In the simulation, we start with an initial

configuration, where a_is are assigned for each site randomly and independently. As time evolves, if neighboring researchers share similar research interests, they will collaborate on a publication. Thus, we start with an ensemble of 10^4 single-author papers. At each discrete time step, we randomly select two nearest-neighboring researchers. If these two researchers share a similar research interest, they will work together and produce a publication of a two-author paper. Within the ensemble, two single-author papers are replaced by one two-author paper. More specifically, the criterion for collaboration is written as

$$|a_i - a_j| < \epsilon_1, \tag{3}$$

where i and j denote two nearest-neighboring sites. The parameter ϵ_1 controls the likelihood of collaboration. If one of the selected researchers has already collaborated with others, the other selected researcher simply joins the collaboration. If both of the selected researchers have already established their own collaborations, a more stringent criterion can be expected for these two research groups to merge into a large collaboration. Such a criterion is parametrized by ϵ_2, where one can expect that $\epsilon_2 \ll \epsilon_1$. In comparison to collaboration between two persons, collaboration between two groups can be more difficult. Such a higher threshold is reflected by the smallness of ϵ_2.

The basic model prescribes a dynamic process of collaboration. The initial configuration consists of 10^4 independent researchers or 10^4 single-author papers on various topics. As time evolves, establishing collaborations among nearby researchers leads to the emergence of multiple-author papers. As a result, the total number of papers decreases accordingly. A typical example is shown in the insert of Fig. 3. As the possible collaborations have been exploited, the process is saturated. Asymptotically, a stable distribution can be achieved by repeatedly applying these operational rules for collaboration. This simple model has only two parameters ϵ_1 and ϵ_2, which provides a convenient way to incorporate the two distinct mechanisms stated in the previous section. The mechanism of parameter λ is different from the mechanism of parameter ϵ. However, it is interesting to note that with a simple prescription of a large value $\epsilon_1=0.4$ for personal collaboration and a small value $\epsilon_2=0.07$ for group collaboration, the 10^4 researchers collaborate to have 3×10^3 papers with a distribution of coauthorship shown in Fig. 3, which is consistent with the empirical data. Without resorting to other complicated issues of collaboration, the overall distribution can be reproduced economically. The personal collaboration parameterized by ϵ_1 is responsible for the quick decay in small n; the group collaboration parametrized by ϵ_2 results in the flattened tail in large n. The depletion of single-author papers can also be reproduced well. In this model, the single-author papers are portrayed as a leftover from the process of establishing collaborations.

IV. DISCUSSIONS

In this paper, we propose simple models for the collaboration among researchers, where two similar but distinct mechanisms are responsible for the personal and group collaborations. It would be interesting to see if these two mechanisms can be further distinguished by other characteristics. *Physical Review Letters* contains research papers from all fields of physics. For different branches of physics, there might be different modes of collaboration. We apply the same analysis to the five journals of *Physical Review*. The results are shown in Fig. 4. The quick decay of small col-

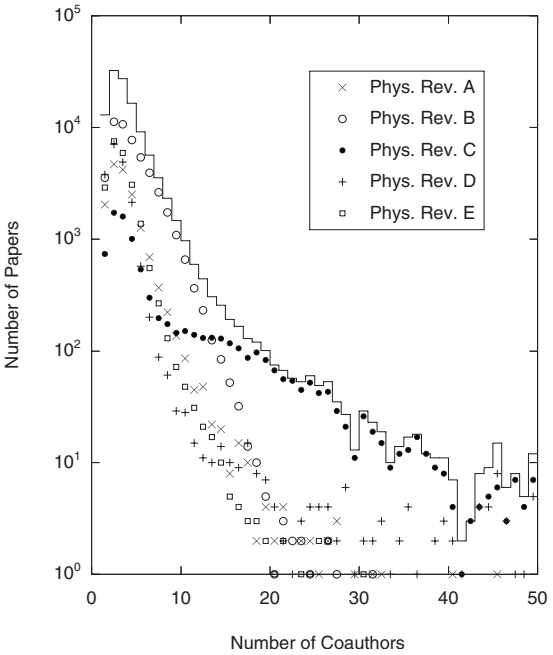

FIG. 4. Distributions of coauthorship in five different journals of *Physical Review*. The data points show the accumulated results in ten years. The incoherent sum of these five journals is shown by the solid histogram, which reproduces the distribution in *Physical Review Letters* as shown in Fig. 1.

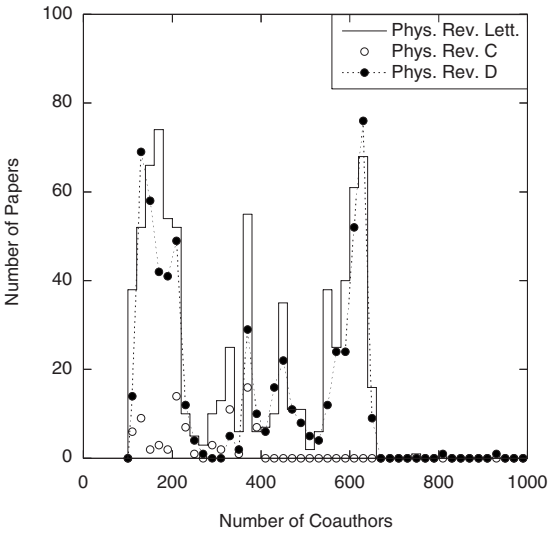

FIG. 5. Distributions of coauthorship in huge collaborations. In this analysis, the maximum number of coauthors 939 appeared on a publication of *Physical Review D* in year 2005. The spectrum of *Physical Review D* (shown by black symbols connected with dotted line) coincides with the spectrum of *Physical Review Letters* (shown by the solid histogram).

laborations and the depletion of single-author papers are the common features. The two-slope structure is absent in *Physical Review B*. The large tail of group collaborations is most significant in *Physical Review C*. It is reasonable to expect that the coverage of research topics in *Physical Review Letters* can be divided into five categories as the *Physical Review* journals. We observe that the incoherent sum from the five journals of *Physical Review* reproduces the distribution from *Physical Review Letters* accurately, see Fig. 4. With this respect, the large tail of group collaborations is basically supported by nuclear physicists in *Physical Review C*; the quick decay of personal collaborations is dominant by condensed-matter physicists in *Physical Review B*.

In this work, we mainly analyze the distribution of coauthors on a published paper up to 100, where a smooth distribution is observed, see Fig. 1. In contrast, the spectrum for very large collaborations cannot be reduced to a continuous distribution, see Fig. 5. Papers with coauthors more than 100 can be found in *Physical Review Letters* (2.5%), *Physical Review C* (1.0%), and *Physical Review D* (3.1%). It is interesting to notice that the spectrum in *Physical Review Letters* matches that in *Physical Review D* quite well. In the continuous distribution shown in Fig. 1, there is no way to identify a specific author. However, the peaks shown in Fig. 5 are easily identified with some of the well-known collaborations. The prominent peak around 200 coauthors can be associated with BELLE and CLEO collaborations. While BABAR and CDF collaborations can be identified with the peak around 600 coauthors [15].

Finally, we present a brief comment on the single-author papers. In this Brief Report, we find that the amount of single-author papers is much fewer than predicted by the Poisson process. In the previous studies based on preprint databases, the single-author papers constitute the largest fraction in the ensemble [11]. Such a difference between journal papers and archive preprints leads to a simple explanation that, compared to multiple-author papers, much more of the single-author preprints cannot find their ways to be published. In this work, the single-author papers in research journals are portrayed as a leftover from the collaboration process. With the rapid advancement of communication technology, developing collaborations among researchers turns out much simpler. The diminishing of single-author papers has been observed for a long time [1]. However, we find that the single-author papers still constitute a sizable fraction of the published papers. In *Physical Review Letters*, 6.8% of papers are single-authored. For other journals, this ratio ranges from 7.1% (*Physical Review B*) to 19.2% (*Physical Review D*). It seems that to work alone is most preferable to high-energy physicists. In contrast, condensed-matter physicists seem more likely to work together. Within the past decade, the ratio of single-author papers presents a noticeable change. Except in *Physical Review C*, the trend toward a decreasing ratio is obvious. However, the trend also indicates that the single-author papers will not totally disappear in the near future.

[1] D. de Solla Price, *Little Science, Big Science* (Columbia University Press, New York, 1963).
[2] S. Arunachalam and M. J. Doss, Curr. Sci. **79**, 621 (2000).
[3] J. P. H. Drenth, J. Am. Med. Assoc. **280**, 219 (1998).
[4] A. M. Diamond, Scientometrics **8**, 315 (1985).
[5] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[6] A. L. Barabási and R. Albert, Science **286**, 509 (1999).
[7] L. D. Costa, M. A. R. Tognetti, and F. N. Silva, Physica A **387**, 6201 (2008).
[8] A. K. Chandra, K. B. Hajra, P. K. Das, and P. Sen, Int. J. Mod. Phys. C **18**, 1157 (2007).
[9] J. J. Ramasco and S. A. Morris, Phys. Rev. E **73**, 016122 (2006).
[10] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).
[11] M. E. J. Newman, Phys. Rev. E **64**, 016131 (2001).
[12] M. E. J. Newman, Phys. Rev. E **64**, 016132 (2001).
[13] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, Physica A **311**, 590 (2002).
[14] In this work, we do not specifically identify each author. In the case that one author publishes more than one paper in the same year, we simply take these papers as independent and from different authors. For the coauthorship networks, this simplification will lead to very different results. For the distribution shown in Fig. 1, however, the results are the same.
[15] BELLE, CLEO, BABAR, and CDF are the names of well-known detectors in high-energy experiments. BELLE collaboration studies CP violation at the KEK B-factory. CLEO collaboration studies beauty and charm quarks at the Cornell Electron Storage Ring (CESR). BABAR collaboration studies B mesons at the Stanford Linear Accelerator Center (SLAC). CDF collaboration studies high-energy particle collisions at the Tevatron (Fermilab).