Some interesting statistics in *Chinese Journal of Physics*

Jiann-wien Hsu* and Ding-wei Huang†

*General Education Center, National Tainan Institute of Nursing, Tai-nan, Taiwan

†Department of Physics, Chung Yuan Christian University, Chung-li, Taiwan

Abstract

We analyze the publications in *Chinese Journal of Physics* over the past two decades. We study the productivity and collaboration of authors in the journal. On average, each author publishes two papers and four coauthors are listed on each paper. However, these numbers are highly biased and not representative for the majority of authors. The productivity distribution can be described by the power law; while the collaboration distribution has an exponential tail. We also obtain analytical formulation to address the correlations between productivity and collaboration.

Introduction

In the traditional classification, human activity is often outside the physics research. With the advance of statistical physics, human activity has attracted lots of attention in physics community in recent years [1, 2]. Econophysics and sociophysics are two typical examples [3]. On one hand, physicists always tend to find out regularity behind the seemingly stochastic behaviors. On the other hand, as the science community grows, such investigations can be interesting in both informetrics and scientometrics.

More than eighty years ago, a simple power law distribution was noticed in scientific productivity [4], which has been often referred to as the Lotka's law. Up to present, there are still many discussions on the validity of this empirical law in various branches of academic activities [5–8]. We think that it can be interesting to provide a simple analysis for the publications in *Chinese Journal of Physics*. In this work, we study the productivity and collaboration of scientists, which result in a publication in *Chinese Journal of Physics*. Correlations between productivity and collaboration have been anticipated [9, 10]. We will first present the empirical data. The analytical formulation will be presented later in the discussions.

Empirical Data

We consider the publications in *Chinese Journal of Physics* over the past twenty years from 1990 to 2009. We sample those articles published by authors with the five most popular Chinese family names: Chen, Lin, Huang, Lee, and Chang. The dataset consists of 563 articles, which are about one-third of the total publications of 1659 articles. As these five common family names cover one-third of the population in Taiwan, such a ratio is not surprising to appear also in scientific productivity measured simply by the number of published papers. We have identified 392 authors. As some of the publications do not provide the authors' first names in full, different authors are further identified by their institutions, research topics, and collaborators. The frequency distribution of their scientific productivity is shown in Fig. 1 with the symbol (•). More than half of the authors, 244 authors, publish only one paper during the twenty years. Only 8 authors publish more than ten papers. The distribution can be described by the Lotka's law, i.e. a power-law decrease

with an exponent of (-2). It is interesting to note that if these authors are identified only by the initials of their first names, the same power law distribution can still be discerned as in Fig. 1 with the symbol $(\times)$. As the ensemble is not large, the two initials for a typical Chinese first name can be used to distinguish these authors effectively. From Fig. 1, each author publishes on average two articles during the past twenty years. However, such an average number is not representative. The distribution is highly non-uniform, where the fluctuations are huge. With the standard notations of statistics, the number of articles per author can be expressed as (2.11 ± 2.27). The huge fluctuations are reflected by a result that the variance is larger than the mean. If a standard were set to publish two articles, three-fourths of the authors shall fall below this standard. In other words, the majority of authors cannot meet the expectation to publish the average number of papers.

We observe that 392 authors publish 563 articles. A simple arithmetic calculation gives a mean of 1.44 articles per author. The discrepancy between these two average numbers, 1.44 and 2.11, is owing to the coauthorship [11]. Judging by these two numbers alone, one would naively expect 50% of overlap in counting the papers, which seems to imply many collaborations among the authors. However, the collaborations shown by the coauthorship are not too frequent. The distribution of coauthorship among authors of these five family names is shown in Fig. 2 with the symbol $(\times)$. The number of coauthors for each paper can be written as (1.47 ± 0.80), which implies the scarce collaboration. Most of the articles, 371 out of 563, show no collaborations. The maximum number of coauthors is six and there are only two such papers. The distribution is exponential, not a power law. When all the other authors are considered, the distribution is shown in Fig. 2 with the symbol $(\bullet)$. The tail part still follows an exponential distribution. The descent has a gentler slope as more coauthors are involved. And the amount of single-author papers is obviously deviated from an exponential distribution. The coauthor number on each paper can be written as (3.69 ± 2.21). Compared to the coauthor numbers among authors of the five family names at (1.47 ± 0.80), the mean and variance increase proportionally. Again, the wide fluctuations imply that a judgement from the arithmetic mean alone can be misleading.

We believe that this dataset is large enough to validate further analysis. First, we subdivide these 563 articles into two groups according to the number of coauthors. There are 307 articles with the coauthor numbers less than or equal to three; and the rest 256 articles have the coauthor numbers larger than three. We have identified 194 authors in the first group

and 255 authors in the second group. There are 57 authors overlapped in these two groups, which is about 15%. The results are shown in Fig. 1 with the symbol (□) for coauthor numbers less than or equal to three, and the symbol (◇) for coauthor numbers larger than three. With intuition, the first group can be associated to the theoretical study and the second group is for experimental study. We find that both groups follow the inverse square law for their scientific productivity.

Second, we subdivide these 392 authors into two groups according to the number of published papers. There are 244 authors who publish only one paper, and the rest 148 authors publish more than one paper. We find that the first group covers 211 papers and the second group covers 439 papers. There are 87 papers overlapped in these two groups, which is about 15%. The coauthor distributions for these two groups are shown in Fig. 2 with the symbol (□) for those published one paper and the symbol (◇) for those published more than one paper. These two distributions follow the same exponential tail. For the small coauthor numbers, the first group has a significant lower weighting. Thus we have an interesting observation that the papers published by these less productive authors have a larger number of coauthors.

Discussions

In this short note, we study some interesting statistics for the publications in *Chinese Journal of Physics*. With simple quantitative measurements, the productivity is represented by the number of papers published by an author; and the collaboration is reflected by the number of coauthors listed on a paper. There are wide fluctuations for both numbers, where the mean and variance are of the same order of magnitude. We show that both distributions are far from the normal distribution. The productivity distribution can be described by the power law; while the collaboration distribution has an exponential tail. With conventional wisdom of arithmetic mean, each author publishes two papers (2.11) and each paper lists four coauthors (3.69). However, these numbers are highly biased and not representative for the majority. If these arithmetic means were taken as standard, we must have a misleading conclusion that most of the authors are unproductive and most of the papers are uncollaborative.

In our analysis, the selected data cover about one-third of the publications in *Chinese*

Journal of Physics. Similar results can be expected for other authors and/or other journals. When the same analysis is applied to *Physical Review Letters* [12], the dataset consists of 642 papers with 211 authors. Their distributions for productivity and collaboration are shown in Figs. 3 and 4, respectively. Similar features can be observed. The productivity for those authors who publish less than twenty papers can be well described by the inverse square law. Deviation to the Lotka's law can be noticed for authors publishing more than twenty papers. Basically such an amazing productivity comes from high-energy experimental groups, which involve hundreds of collaborators. The high productivity is strongly correlated to the huge collaborations, which obviously cannot be described by the Lotka's law [13]. Among the 642 papers, there are 360 papers listing more than one hundred coauthors. With these huge collaborations excluded, the distribution shown in Fig. 4 is basically the same as shown in Fig. 2. We note that the power laws shown in Fig. 1 and Fig. 3 have the same exponent; while the exponential tails shown in Fig. 2 and Fig. 4 have slightly different slopes.

We study the productivity distribution shown in Fig. 1 and the collaboration distribution shown in Fig. 2. These distributions reflect two different aspects of scientific activity. However, correlations between these two aspects can be expected. Consider N authors to publish M papers. Presume the productivity and collaboration follow respectively the power law and the exponential trend. Then we have the following constraints,

$$\sum_{i=1} \frac{a_0}{i^2} = N \quad , \tag{1}$$

$$\sum_{i=1} b_0 \; e^{-c(i-1)} = M \quad , \tag{2}$$

where a_0 denotes the number of authors who publish only one paper, b_0 denotes the number of papers which list only one author, and c denotes the slope of exponential tail. It is interesting to note that both a_0 and b_0 are not free parameters. With simple algebra, we obtain the solutions

$$a_0 \; = \; \frac{6}{\pi^2} \; N \sim 0.61 \; N \quad , \tag{3}$$

$$b_0 \; = \; \frac{e^c - 1}{e^c} \; M \sim 0.63 \; M \quad (\text{if} \quad c = 1) \quad . \tag{4}$$

The two average numbers can also be expressed as follows,

$$a \; = \; \frac{a_0}{N} \; \sum_{i=1}^{I_a} \frac{1}{i} = \frac{6}{\pi^2} \; \sum_{i=1}^{I_a} \frac{1}{i} \quad , \tag{5}$$

$$b \;=\; \frac{b_0\; e^c}{M}\;\sum_{i=1} i\; e^{-ci} = \frac{b_0\; e^{2c}}{M(e^c-1)^2} = \frac{e^c}{e^c-1} \sim 1.58 \;\;(\text{if}\;\; c=1)\;\;, \tag{6}$$

where a denotes the average number of papers published by each author and b denotes the average number of coauthors listed on each paper. We introduce a cutoff I_a in Eq. (5) to avoid divergence in the summation. This cutoff has a physical interpretation as the number of papers published by the most productive author. A reasonable setting can be

$$\frac{a_0}{(I_a)^2} \sim 1 \quad \text{or} \quad I_a \sim \sqrt{a_0} \sim 0.78\;\sqrt{N}\;\;. \tag{7}$$

The average number a is sensitive to the cutoff I_a, which can be related to the observed feature that the arithmetic mean is determined by the extreme minority and does not reflect the majority. In contrast, there is no need of cutoff to obtain the other average number b, which is determined solely by the slope c. The correlation between productivity and collaboration implies the following constraint between the two average numbers,

$$N \cdot a = M \cdot b\;\;. \tag{8}$$

Without the parameter c, N and M are related by the constraint. If we want to treat N and M as independent from each other, the introduction of parameter c is necessary. Otherwise, the constraint of Eq. (8) cannot be satisfied. For the data from *Chinese Journal of Physics*, $(N, M) = (392, 563)$. The analytic analysis gives $(a_0, a, I_a) = (239, 2.02, 15)$ and $(b_0, b, c) = (399, 1.41, 1.23)$. The distributions shown in Figs. 1 and 2 can be reproduced. It can be interesting to further investigate the validity of this analytic analysis in other situations.

Acknowledgement

We thank Dr. Fisenko to bring this problem to our attention by sending us their recent work of Ref. [5].

[1] C. Castellano, S. Fortunato, and V. Loreto, *Rev. Mod. Phys.* **81** (2009) 591.

[2] V. M. Yakovenko and J. B. Rosser,, *Rev. Mod. Phys.* **81** (2009) 1703.

[3] For a review, see *Econophysics and Sociophysics: Trends and Perspectives*, edited by B. K. Chakrabarti, A. Chakraborti, and A. Chatterjee (Wiley VCH, Weinheim 2006).

[4] A. Lotka, *J. Washington Acad. Sci.* **16** (1926) 317.

[5] N. V. Pavlyukevich, O. G. Penyazkov, and S. P. Fisenko, *Journal of Engineering Physics and Thermophysics* **82** (2009) 608.

[6] M. Petek, *Scientometrics* **75** (2008) 175.

[7] I. Rowlands, *Aslib Proceedings* **57** (2005) 5.

[8] J. Baker, J. Robertson-Wilson, and W. Sedgwick, *Journal of Sport and Exercise Psychology* **25** (2003) 477.

[9] A. Akakandelwa, *African Journal of Library Archives and Information Science* **19** (2009) 13.

[10] B. M. Gupta and C. R. Karisiddippa, *Scientometrics* **44** (1999) 129.

[11] J. W. Hsu and D. W. Huang, *Phys. Rev.* **E80** (2009) 057101.

[12] To obtain a similar size of ensemble and a reasonable comparison, we consider only the papers with an address in Taiwan.

[13] H. Kretschmer and R. Rousseau, *Journal of the American Society for Information Science and Technology* **52** (2001) 610.

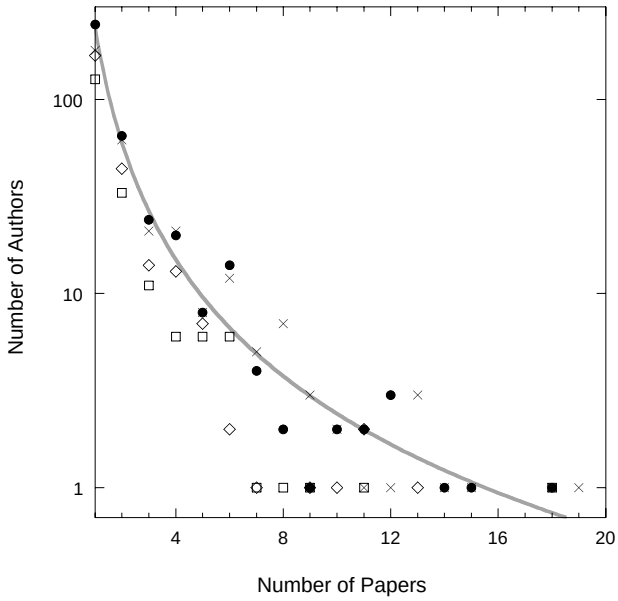

FIG. 1: Productivity distribution in *Chin. J. Phys.* The grey line shows the power law with exponent (-2).

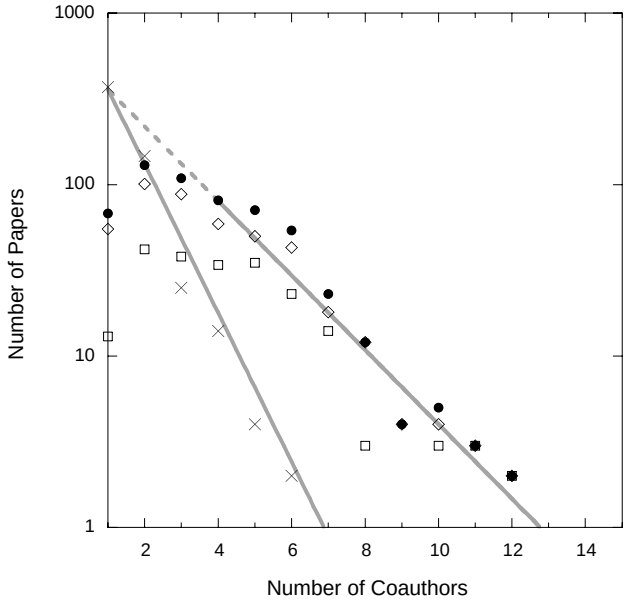

FIG. 2: Collaboration distribution in *Chin. J. Phys.* The grey lines show the exponential decrease with slopes (-1) and (-0.5).

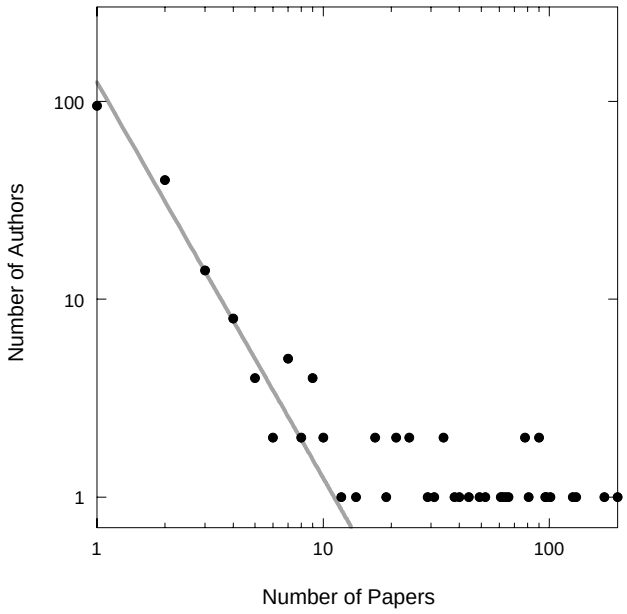

FIG. 3: Productivity distribution in *Phys. Rev. Lett.* The grey line shows the power law with exponent (-2), which is the same as shown in Fig. 1.

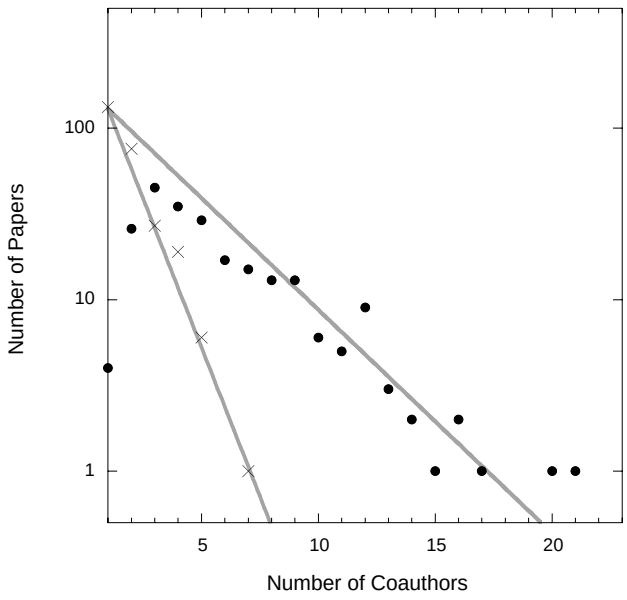

FIG. 4: Collaboration distribution in *Phys. Rev. Lett.* The grey lines show the exponential decrease with slopes (-0.8) and (-0.3), which are different from those shown in Fig. 2.