Comparative Analysis of Author Productivity of Different Domains: in Consideration of the Effect of Sample Size Dependency of the Statistical Measures

Fuyuki Yoshikane
Graduate School of Education, University of Tokyo
7–3–1 Hongo, Bunkyo-ku, Tokyo, Japan
E-Mail: fuyuki@p.u-tokyo.ac.jp

Abstract

We analyze and compare the characteristics of the author productivity in academic conference papers of four different domains, focusing the degree of concentration in the distributions. In this analysis, we pay attention to a peculiar feature of author productivity data, i.e. most of the statistical measures change systematically according to changes in the sample size, which causes difficulty in comparing different data of different size. One way to compare author productivity in different domains is analyzing the developmental profiles of measures for growing sample size. This approach is applied in this study.

1 Introduction

The survey and analysis of actual states concerning such a bibliometric phenomenon as publication of journals or performance of researchers, which is based on bibliographical information sources, e.g. bibliographies or indexes, has been one of the main themes in bibliometrics. Various measures which are used in the analysis have been proposed, for various purposes, from various viewpoints (cf. Kishida 1996).

For instance, as a basic index to measure the degree of concentration in distribution, Gini's index G, or Pratt's measure C (Pratt 1977) is well-known (cf. Egghe & Rousseau 1990). However, almost all statistical measures, inclusive of the two measures mentioned above, change systematically according to the change of the sample size when the sample is statistically insufficient. Then, we cannot compare characteristics of the populations directly by those measures in the survey based on sampling. And, statistically speaking, most of the bibliometric samples are insufficient as a sample because of the existence of unseen events.

If we are to pay attention to the effect of the sample size dependency of the statistical measures, we must either use some sample size invariant measures or analyze the developmental profiles of measures for growing sample size, in order to compare the different samples of different size. In this study, by the latter approach, we analyze the degree of concentration of the author productivity in academic conference papers of different domains. This study will provide a framework which is applicable to the analysis of bibliometric phenomena based on samples of the statistically peculiar characteristics.

2 Background

2.1 The Concept of Concentration: Dispersion and Inequality

The bibliometric laws show ‘concentration’ in distribution. For example, Lotka’s law (Lotka 1926) shows that a large number of papers of some domain are concentrated in particularly productive authors, and Bradford’s law (Bradford 1934) shows that most of the papers of some topic are concentrated in a few core journals. These laws show not only that most items, e.g. papers of some domain, are concentrated in a few sources, e.g. authors, but also that the other items disperse over many sources.

Many measures of concentration have been already proposed. However, the concept of ‘concentration’ is not clearly defined in the proposed measures. Concerning the proposed measures, Egghe & Rousseau (1991) pointed out the problem that the measures are thrown into a typical circular reasoning: “the notion of concentration is defined through the value of a measure used to measure concentration.”

Concerning concentration, generally speaking, we have two concerns. One is to observe how many sources items disperse over, i.e. the absolute scale of dispersion. The other is to observe inequality among those sources. We regard distributions where items disperse over more sources as less concentrated, and regard more highly skewed distributions as more concentrated. Ray & Singer (1973) named these two viewpoints ‘absolute concentration’ and ‘relative concentration’ respectively[1] .

Some existing concentration measures are sensitive to both dispersion (absolute concentration) and inequality (relative concentration), and some are sensitive only to inequality. Yoshikane (2000) analyzed existing measures and classified them from the above two viewpoints. In this study we distinguish between absolute and relative concentration, and use measures, considering their sensitivity to both types of concentration.

2.2 Sample Size Dependency of Statistical Measures

Mathematically, most bibliometric data are insufficient as a sample, because it is statistically expected that not all the events (sources) in the population appear in the sample data. For instance, in most author productivity data, all the authors in the domain are not exhausted. There are unseen authors. The coefficient of loss is a convenient measure which shows to what extent the data is insufficient, by giving the ratio of loss of the estimated number of authors calculated by the sample relative frequencies as the estimates of population probabilities against the empirical number of authors in the original sample (Chitashvili & Baayen 1993):

$$\begin{aligned} C_L &= (V(N) - \hat{E}[V(N)])/V(N) \\ &= \frac{\sum_{m\geq 1} V(m,N)(1 - p(i_{[f(i,N)=m]}, N))^N}{V(N)} \end{aligned} \qquad (1)$$

[1] It seems that there is a confusion between ‘absolute concentration’ and ‘relative concentration’ in discussing concentration. For example, Pratt (1977) uses the term ‘relative concentration’ as the relative degree of concentration of one source against the whole distribution. In this study, however, we use this term as the relative inequality between sources (authors) according to Ray & Singer.

where $f(i,N)$ represents the frequency of an author a_i, $p(i,N) = f(i,N)/N$, i.e. sample relative frequency, $V(N)$ represents the number of authors, $V(m,N)$ represents the number of authors appearing m times, and N is the sample size.

When the coefficient of loss is large, most statistical measures crucially depend on the sample size (Tweedie & Baayen 1998), which makes proper comparison and interpretation of different samples of different sizes impossible. One way to compare author productivity in different domains, whose sample data most probably varies in size, is analyzing the developmental profiles of measures for growing sample size. Another way is to look for sample size invariant measures. In the previous study (Yoshikane & Kageura 1999), we showed the usefulness of the latter approach. In this paper, we examine the potential of the former approach.

3 Methodology

3.1 The Data

Here, we explain the data used in this analysis. We used a bibliographic database of academic conference papers, provided by National Institute of Informatics, Japan. From the database, we extracted the records of conferences hosted by four different academic societies, i.e. 'Institute of Electrical Engineers of Japan,' 'Japan Society for Bioscience, Biotechnology, and Agrochemistry,' 'Architecture Institute of Japan' and 'Information Processing Society of Japan' from 1992 to 1997. The author-paper relation in the data is regarded as an indication of author productivity.

In the study of author-paper relation, the problem associated with multiple authorship arises. In this paper, we credit each collaborating author with a full contribution. So, in the following statistical arguments, the total number of author tokens, instead of the number of papers, is regarded as the sample size[2] . Thus the author productivity addressed in this study indicates the degree of researchers' activity in an abstract sense, as represented by the occurrence of names in the conference papers, rather than the accurate evaluation of their contributions.

Table 1 shows the basic quantities of each sample, i.e. the sample size (N_0), the number of authors in a sample of size N_0 ($V(N_0)$) and the coefficient of loss C_L. Henceforth for succinctness, we call four samples by their 'domains', namely electrical engineering, biochemistry, architecture, information processing.

Data	N_0	$V(N_0)$	C_L
Electrical Engineering	75685	25230	0.241
Biochemistry	71974	21315	0.229
Architecture	166941	27143	0.159
Information Processing	79372	24271	0.225

Table 1: The basic quantities of the data for four domains

In all the domains, the coefficient of loss exceeds 0.15, which means that the number of author is underestimated by more than 15%, if the population probabilities are estimated by the sample relative frequencies. As mentioned above, most statistical measures crucially depend on the sample size when the coefficient of loss is large. Therefore we cannot compare the characteristics of the

[2] For convenience, we use 'papers' and 'author tokens' interchangeably.

populations directly on the basis of the values of the measures which are calculated by the samples themselves.

3.2 Concentration Measures

From the two viewpoints, i.e. the scale of dispersion (absolute concentration) and the degree of inequality (relative concentration), we evaluate concentration in author productivity distributions. We use the number of authors (V) and Gini's index (G) to measure concentration. We do not have to integrate the evaluation of absolute concentration and that of relative concentration into one measure, because our aim is not to determine the ranking of domains in the degree of concentration of the author productivity but to describe the characteristics of domains.

Absolute concentration is measured by observing how many authors papers disperse over. As the measure of relative concentration, on the other hand, we select Gini's index.

$$G(N) = \sum_{i=1}^{V(N)} \sum_{j=1}^{V(N)} \frac{|f(i,N) - f(j,N)|}{2\mu V(N)^2} \tag{2}$$

where μ represents the mean frequency. There are two reasons why we select Gini's index. Firstly, G is insensitive to the number of authors, that is, absolute concentration. Secondly, other many concentration measures, such as HH (Herfindahl 1950), CCI (Horvath 1970) and CON (Ray & Singer 1973), have an undesirable characteristic in that they are extremely sensitive to the most productive authors, but G does not (Yoshikane 2000).

3.3 Developmental Profiles of the Measures

As shown above, the coefficient of loss is large in our samples. Therefore, it is anticipated that statistical measures, including V and G, change systematically according to the change of the sample size. When we compare the different samples of different size on the basis of sample size dependent measures, we have to take the dynamism of the measures into consideration.

In order to observe the change of V and G, we carry out the random Monte Carlo sub-sampling of 1000 trials for 20 equally-spaced intervals and calculate the values of the measures for each sample size. By tracing the developmental profiles of the measures, we are able to describe the characteristics of the domain itself beyond the given sample size.

4 Analysis

4.1 Comparison of the Original Samples

Figure 1 shows the number of authors V and Gini's index G of the original sample of each domain. V and G represent the scale of dispersion (absolute concentration) and the degree of inequality (relative concentration), respectively. Among the four domains, architecture has the highest value both in V and in G. As for the remaining domains, i.e. electrical engineering, biochemistry and information processing, a negative correlation between the two measures is observed. That is to say, in the domain whose papers are concentrated on fewer authors, the papers are more concentrated on especially productive authors among them.

It is possible to compare the samples on the basis of the values shown in figure 1. However, we must pay attention to the fact that the result is no more than of the comparison of the original samples themselves. That is to say, the object of the above mentioned analysis is the data of

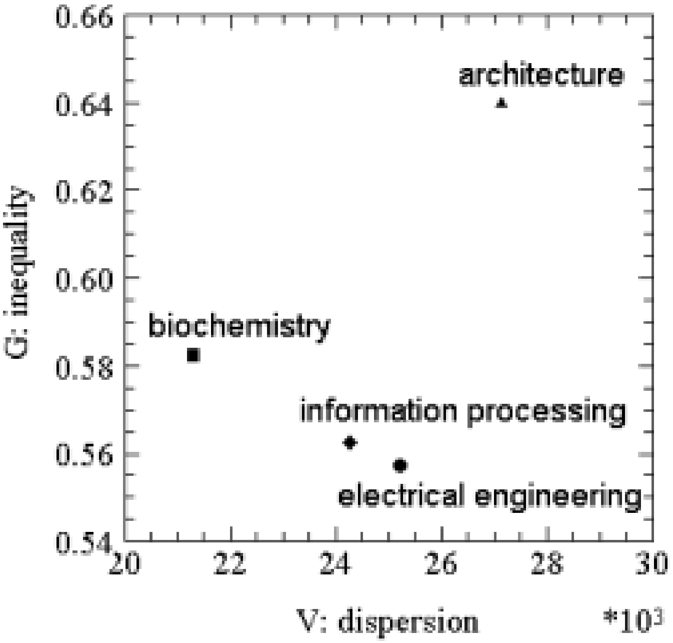


Figure 1: Comparison of the four original samples by V and G

the papers from 1992 to 1997 in each academic conference. Therefore we cannot generalize the characteristics of the whole of each domain, from the result of the comparison based on the limited samples.

4.2 Dynamism of the Measures

Figure 2 and 3 plot the developmental profiles of the two measures obtained by the random Monte Carlo sub-sampling. The figures show that both $V(N)$ and $G(N)$ increase systematically according to growth of the sample size N in all the domains. The larger the sample size grows, the lower the degree of absolute concentration becomes and the higher the degree of relative concentration becomes. The figures imply that, when we compare the different samples of different size, the result of comparison depends on the sample size of each domain unfortunately.

It is theoretically obvious that the number of authors $V(N)$ increases systematically when the sample size N is increased. Based on the assumption that the authors are binomially distributed, and assuming that there are S authors in the population, to each of which the population probability p_i is assigned, we can obtain the expected number of authors in the sample of size N by using the following equation.

$$E[V(N)] = \sum_{m=1}^{N} \sum_{i=1}^{S} \binom{N}{m} p_i^m (1-p_i)^{N-m} \tag{3}$$

This formula shows that we need a sample of a very large size if we are to obtain a statistically sufficient sample, when the number of authors in the population is relatively large and the population probabilities of the authors are generally low.

Comparing the four domains at the same sample size, we can observe that electrical engineering has the highest values of $V(N)$. And electrical engineering is followed by information processing, biochemistry and architecture. As for $G(N)$, biochemistry has the highest values and architecture follow it. Electric engineering and information processing have the lowest values. Judging from the correspondence between the locus of $G(N)$ in electric engineering and that in information processing, we can say that the two domains have very similar characteristics in absolute concentration.

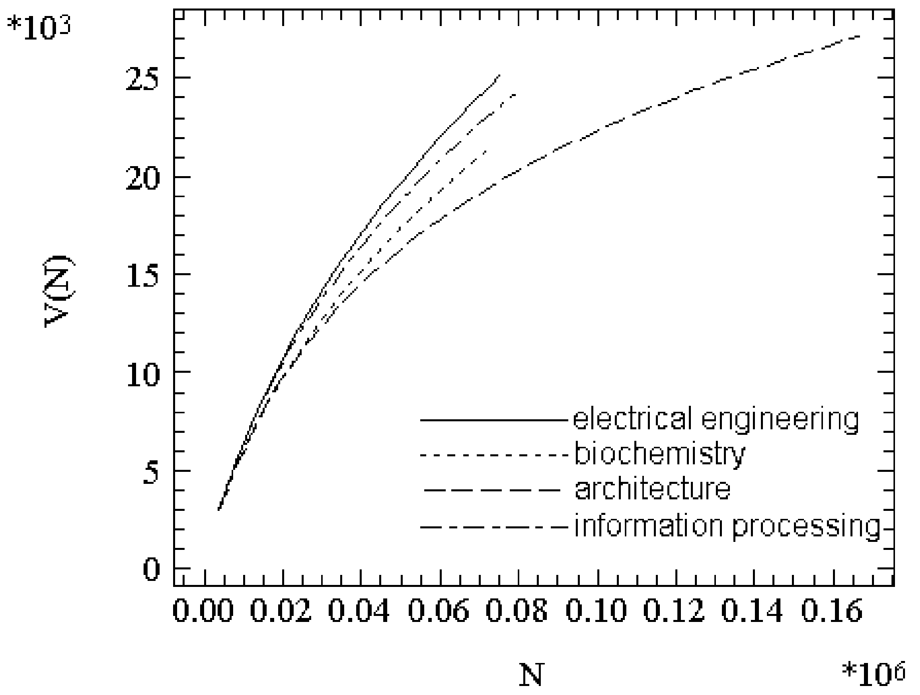


Figure 2: Change of the values of the number of authors $V(N)$

4.3 Characteristics of Each Domain

In order to compare the productivity of authors in the four domains whose sample data vary in size, we have plotted the developmental profiles of $V(N)$ and $G(N)$. We further visualize the pattern of characteristics at the same sample size in figure 4. Figure 4 shows four snapshots of the pattern at the sample size $N = 10000, 30000, 50000, 70000$.

In figure 4, the correlation between the two measures is not observed. For example, biochemistry has the highest value of $G(N)$ whatever sample size we take. As for $V(N)$, however, it is located in the center among the four domains (especially at the large sample sizes). That is, the domain of high relative concentration is not necessarily high in absolute concentration.

The pattern of the four domains changes according to growth of the sample size. At the early stage, electrical engineering, architecture and information processing show similar characteristics while biochemistry is isolated. However, architecture deviates from electrical engineering and information processing greatly when the sample size increases. At the large sample size, as compared with electrical engineering and information processing, architecture has considerably low value of $V(N)$ (and slightly high value of $G(N)$). That is, when the number of papers is large, those papers are produced by a relatively small number of authors in architecture.

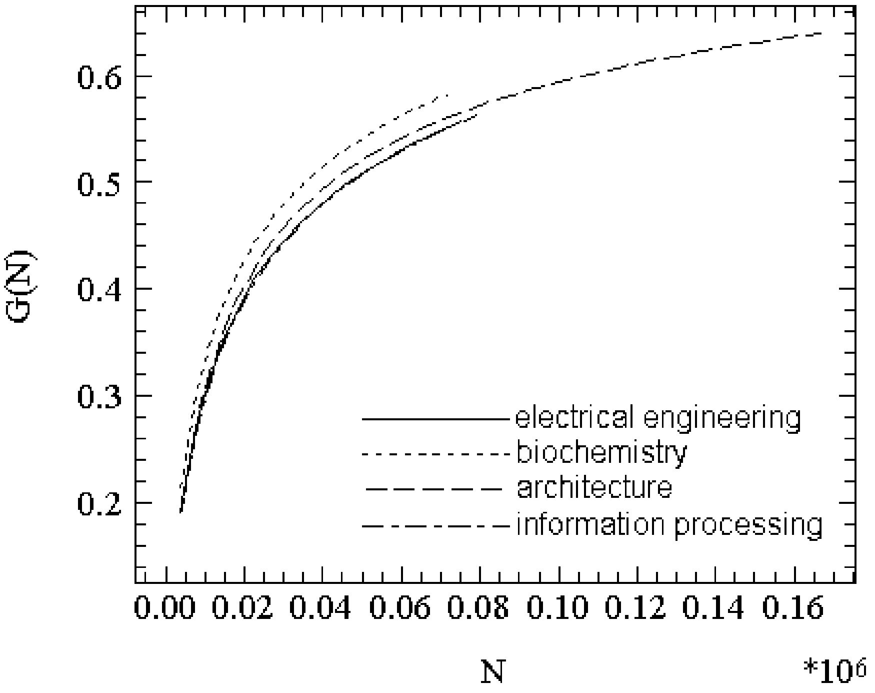


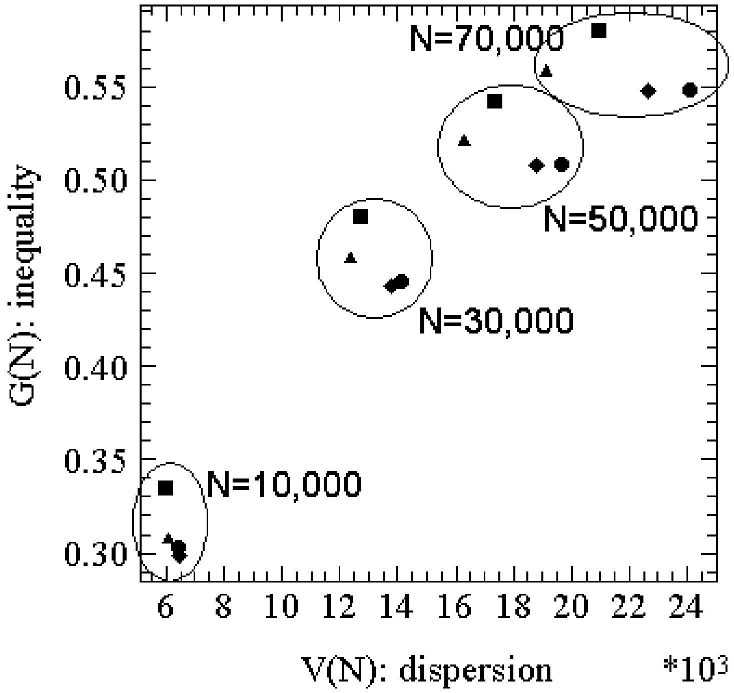
Figure 3: Change of the values of Gini's index $G(N)$

The characteristics of each domain can be summarized as follows. Biochemistry shows higher relative concentration (higher values of $G(N)$) than the other three domains: inequality between productive authors and sterile authors is big in this domain. Architecture shows higher absolute concentration (lower values of $V(N)$): the papers in this domain can be covered by a small number of authors. This feature appears notably when the number of papers is large. Electrical engineering and information processing show similar characteristics. The two domains are low both in absolute concentration and in relative concentration. As the number of papers increases, however, absolute concentration in information processing becomes higher than that in electrical engineering gradually.

5 Conclusions

Considering the dynamism of the measures, we have compared the characteristics of the four domains. The result does not correspond with the features observed in the original samples themselves, which is clearly shown in figure 1 and 4. For example, architecture exceeds the other domains in both measures in the comparison of the original samples. However, it is caused only by the sample size of architecture which is twice as large as those of the other domains. At the same sample size, architecture has lower value of $V(N)$ than electrical engineering and information processing, and lower values of $G(N)$ than biochemistry. Like this, it is not possible to regard the features observed in a sample as the characteristics of a domain itself.

This study and the previous study examined two methods to compare different data of different size, i.e. analyzing the developmental profiles of measures for growing sample size, and the use of sample size invariant measures. As the next step, we are going to integrate the interpretation of the results obtained by both methods, by which the many-sided characteristics of author productivity



circle: electrical engineering, square: biochemistry,
triangle: architecture, diamond: information processing

Figure 4: Comparison of the four domains by $V(N)$ and $G(N)$

can be properly described.

References

Bradford, S. C. (1934) “Sources of information on specific subjects.” *Engineering*, 137, p. 85–86.

Chitashvili, R. J. and Baayen, R. H. (1993) “Word frequency distributions.” In: Hrebicek, L. and Altmann, G. (eds.) *Quantitative Text Analysis.* Trier: Wissenschaftlicher Verlag. p. 54–135.

Egghe, L. and Rousseau, R. (1990) “Elements of concentration theory.” In: Egghe, L. and Rousseau, R. (eds.) *Informetrics*, 89/90. Amsterdam: Elsevier. p. 97–137.

Egghe, L. and Rousseau, R. (1991) “Transfer principles and a classification of concentration measures.” *Journal of the American Society for Information Science*, 42(7), p. 479–489.

Herfindahl, O. C. (1950) *Concentration in the Steel Industry*, Ph.D. dissertation, Columbia University.

Horvath, J. (1970) “Suggestion for a comprehensive measure of concentration.” *The Southern Economic Journal*, 36, p. 446–452.

Kishida, K. (1996) “Some characteristics of scientometric indicators.” *Journal of Japan Indexers Association*, 20(2), p. 1–11.

Lotka, A. J. (1926) “The frequency distribution of scientific productivity.” *Journal of the Washington Academy of Sciences*, 16(12), p. 317–323.

Pratt, A. D. (1977) "A measure of class concentration in bibliometrics." *Journal of the American Society for Information Science*, 28(5), p. 285–292.

Ray, J. L. and Singer, J. D. (1973) "Measuring the concentration of power in the international system." *Sociological Methods & Research*, 1(4), p. 403–437.

Tweedie, F. J. and Baayen, R. H. (1998) "How variable may a constant be?: Measures of lexical richness in perspective." *Computers and the Humanities*, 32, p. 323–352

Yoshikane, F. (2000) "Concentration in bibliometric distributions: The notion of concentration and concentration measures." *Journal of Japan Society of Library and Information Science*, 46(1), p. 18–32.

Yoshikane, F. and Kageura, K. (1999) "On the potential of sample size invariant measures for the comparative analysis of author productivity data." *7th Conference of the International Society for Scientometrics and Informetrics*, p. 547–557.