Comparative Analysis of Co-authorship Networks of Different Domains: the Growth and Change of Networks

Fuyuki Yoshikane
Faculty of University Evaluation and Research, National Institution for Academic Degrees, 3-29-1 Otsuka, Bunkyo-ku, Tokyo, 112-0012 Japan. E-Mail: fuyuki@niad.ac.jp

Kyo Kageura
National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan. E-mail: kyo@nii.ac.jp

Abstract
There are many studies which try to analyze 'co-authorship networks' and to describe their patterns. However, most studies describe mainly static networks, and they do not take into account the statistical peculiarity of author productivity data, i.e., the sample size dependency of statistical measures. In this study, we turn our attention to the change of co-authorship networks according to increase in the sample size. We consider 'the dynamic characteristics' of networks, and try to compare co-authorship networks of different domains.

1. Introduction

In recent years, academic researches have become more complex, and people regard research collaboration as more and more important. We may see it, for example, through the fact that in many fields (especially in natural sciences) both the number of coauthors and the ratio of coauthored papers are increasing (e.g., Devilliers, 1984; Drenth, 1988; O'Neill, 1988; Yitzhaki & Ben-Tamar, 1990). Thus, we can say that it is an important issue to arrange systems, such as grants-in-aid, for supporting research collaboration (e.g., intercollegiate collaboration and collaboration with industrial circles). For this issue, first of all, we have to grasp the present situation of research collaboration.

The situation of research collaboration can be measured to some extent through analyzing the products of collaboration, i.e., coauthored papers. There are many studies which try to analyze 'co-authorship networks' and to describe their

patterns[1]. For instance, some studies proposed indices for measuring the link strength between nodes (authors) and analyzed actual co-authorship networks using those indices (e.g., Narin et al., 1991; Arunachalam et al., 1994; Kretschmer & Gupta, 1998; Kundra & Kretschmer, 1999), and others applied statistical methods such as Factorial Correspondence Analysis (FCA) to the analysis of co-authorship networks (e.g., Elalami et al., 1992; Okubo et al., 1992). However, most studies describe mainly static networks, and they do not take into account the statistical peculiarity of author productivity data, i.e., the sample size dependency of statistical measures. In this study, we consider the change of co-authorship networks. We consider 'the dynamic characteristics' of networks, and try to compare co-authorship networks of different domains.

This study analyzes the change in the mass and variety of each author's 'ego-centered networks' (Wasserman & Faust, 1994) according to increase in the number of papers, as a start point of dealing with the growth of sociocentric networks of co-authorship.

2. Selection of Measures

In this study, we examine (i) the growth of the mass of relations in networks by observing the number of partners (links) of each author (node), and examine (ii) the change of the variety of relations in networks by observing inequality of relationship strength (link strength) with each partner. We use the number of partners (V) as an index for the first viewpoint, and used Gini's index (G) as an index for the second viewpoint. G is obtained by:

$$G = \sum_{i=1}^{V} \sum_{j=1}^{V} \frac{| f_i - f_j |}{2\mu V^2}$$

where V represents the number of partners, f_i represents the frequency of a partner a_i, and μ represents the mean frequency.

Many indices for measuring inequality have been proposed. Among them, we selected G for the following two reasons. One reason is that G is insensitive to the number of partners (Yoshikane, 2000). This feature is desirable for our aim, because we wish to observe, as the 'variety' of relations in networks, pure inequality where the influence of the number of partners is removed. (We also observe the number of partners, V, as the 'mass' of relations.) The other reason is that G is sensitive to all

---

[1] Social network analysis studies, including co-authorship network studies, are reviewed in Otte & Rousseau (2002).

partners equally (Yoshikane, 2000). This feature is also desirable, because we wish to observe inequality on the whole, not to attach much importance to the relationship strength with particular partners (e.g., the most frequent partners).

3. The Data and Target of Analysis

This study analyzes not 'diachronic' but 'synchronic' dynamics. The reason why we deal synchronic dynamics before diachronic ones is due to the data peculiarity mentioned below.

The data used in this study were extracted from a bibliographic database of academic conference papers, provided by the National Institute of Informatics, Japan. From the database, we extracted the records of conferences hosted between 1992 and 1997 by four different academic societies: the Institute of Electrical Engineers of Japan; the Information Processing Society of Japan; Society of Polymer Science, Japan; and the Japan Society for Bioscience, Biotechnology, and Agrochemistry. We regard these as the data sampled from the whole author productivity data in each of the four domains, i.e., electrical engineering, information processing, polymer science, and biochemistry. Table 1 shows the basic quantities of each domain: the number of papers N, the number of authors A(N), and the coefficient of loss C_L.

C_L, is a convenient measure for checking the reliability of data as a sample. C_L calculates the difference between the actual number of authors who appear in a sample of size N and the expected number of authors estimated by using the sample relative frequencies at a given size N as estimates of the population probabilities (Chitashvili & Baayen, 1993):

$$C_L = \frac{A(N) - \hat{E}[A(N)]}{A(N)} = \frac{\sum_{m \geq 1} A(m,N)\left(1 - p\left(i_{[f(i,N)=m]}, N\right)\right)^N}{A(N)}$$

where f(i,N) and p(i,N) represent the frequency and the sample relative frequency of an author a_i in a sample of size N respectively, and A(m,N) represents the number of authors appearing m times. In all the domains C_L exceeds 0.2, which means that the number of authors is underestimated by more than 20% if the population probabilities are estimated by the sample relative frequencies. It is known that, when C_L is large, not only the number of authors A(N) but also most of the statistical measures calculated by frequencies of events (e.g., authors) crucially depend on sample size N (Tweedie & Baayen, 1998). In most author productivity data, as well as in our data, there are many

low-frequency authors, which makes C_L large (Kageura, 1999).

Likewise, C_L is large in 'the co-authorship statistics' of our data. Table 2 shows the average values per author of the number of partners $V_{av}(N)$, those of Gini's index $G_{av}(N)$, and those of the coefficient of loss C_{Lav}. We calculated C_{Lav} on the basis of the partner frequency distribution of each author.[2] From the fact that C_{Lav} is large in all the domains, both $V_{av}(N)$ and $G_{av}(N)$ are expected to depend on N.

The four domains vary in sample size. Therefore, we cannot directly compare their populations (not their samples) on the basis of sample size dependent measures. Of course, we cannot compare their time-series transitions without considering the sample size dependency. Thus, in this study, we observe the transitions of $V_{av}(N)$ and $G_{av}(N)$ according to increase in the sample size, instead of their time-series transitions. That is, this study regards the sample size as an exogenous factor, and compares the four domains from the viewpoint of the co-authorship patterns conditioned by the number of published papers.

4. Analysis

In order to observe the growth and change of networks, we carried out Monte-Carlo simulation in which we perform 1000 random sub-samplings for each 1000 interval of the sample size (i.e., N=1000, 2000, 3000, ...). For each sample size, we calculated the average value of 1000 trials for each of the two measures, the number of partners $V_{av}(N)$ and Gini's index $G_{av}(N)$ (i.e., the average values per author of V(N) and G(N)). Moreover, we plotted the growth rates of these measures at each sample size.

Figures 1a and 1b show the developmental profiles of $V_{av}(N)$ and $G_{av}(N)$, which illustrate the transitions of the measures according to changes in the sample size[3]. On the other hand, Figures 2a and 2b show the growth rates of them. By normalizing the sample size (the number of papers), we compare the four domains with the same condition.

$V_{av}(N)$: the number of collaborating partners

---

[2] As for isolated authors who have no coauthored papers, C_L cannot be calculated by the formula described previously. However, we do not necessarily need C_L of all authors for claiming the unreliability of the data. That is, if there are some authors who have large C_L, the co-authorship data is not reliable. So, when we calculated C_{Lav}, isolated authors were excluded.

[3] In figure 1b, $G_{av}(N)$, which shows 'inequality', increases according to growth in the sample size, because we drop zero sources (unseen authors who do not appear in a sample). If we keep them in calculating $G_{av}(N)$, we may have a different result (see Egghe, 2002).

Comparing the four domains by $V_{av}(N)$ at the same sample size, we can observe that biochemistry has the highest values, and that electrical engineering and polymer science are following it while information processing has considerably low values. That is to say, on the condition that the domains are equal in the number of papers (or places to publish papers), researchers in biochemistry require more collaborating partners than do those in the other domains. On the other hand, researchers in information processing collaborate with the least partners.

Roughly speaking, this order corresponds to the order of the number of coauthors per paper. In Table 3, we show the number of coauthors per paper C in the original data for each domain.[4] We can easily imagine that, if the number of coauthors per paper is large, the number of partners per author also becomes large.

However, observing the transition of $V_{av}(N)$, we find that the order is not fixed but dependent on the sample size. In Figure 1a, it is observed that $V_{av}(N)$ in polymer science becomes smaller than that in electric engineering around N=7000. That is, when the number of papers is small, the number of partners per author in polymer science is larger than that in electrical engineering although polymer science is smaller than electrical engineering in the number of coauthors per paper.

We can observe the decline of the growth of $V_{av}(N)$ in polymer science more clearly in Figure 2a. According to increase in the sample size, the growth rate of $V_{av}(N)$ in polymer science declines sharply. When the sample size exceeds 20000, the growth rate in polymer science becomes almost equal to that in information processing, which has the lowest values of $V_{av}(N)$ among the four domains.

From the growth rate of $V_{av}(N)$ in polymer science in this simulation, we can assume that, in this domain, researchers require a relatively large number of partners even for publishing a small number of papers, but, by collaborating with these partners many times, they can publish many paper without searching many new partners.

$G_{av}(N)$: the inequality of collaborating frequencies among the partners

From Figure 1b, it is observed that polymer science has the highest values of $G_{av}(N)$. Polymer science is followed by biochemistry, electrical engineering, and information processing. In this figure, similar domains show similar characteristics, i.e., the two domains related to chemistry (polymer science and biochemistry) are high in $G_{av}(N)$ while the two domains related to engineering (electrical engineering and

---

[4] C is not a measure calculated by frequencies of events. So, C is free from the problem of the sample size dependency. (Of course, it is not free from the problem of the confidential interval. In a small sample, the interval of the possible error must be large.)

information processing) are low in $G_{av}(N)$. This is also observed in Figure 2b, which shows the growth rate of $G_{av}(N)$.

As for these four domains, we can say that a researcher engaged in chemistry collaborates with their partners with various frequencies, and that a researcher engaged in engineering collaborates with their partners with relatively similar frequencies.

Summary of the results

The characteristics of each domain can be summarized as follows. As for information processing, both $V_{av}(N)$ and $G_{av}(N)$ are low: in this domain researchers are collaborating with a relatively small number of partners with relatively similar frequencies. As for biochemistry, both $V(N)$ and $G_{av}(N)$ are high: in this domain researchers are collaborating with a relatively large number of partners with various frequencies.

Electrical engineering and polymer science are medium in the number of partners $V_{av}(N)$. As for $G_{av}(N)$, polymer science has the highest values among the four domains: in this domain researchers have both very frequently collaborating partners and very rarely collaborating partners. On the other hand, electrical engineering shows similar characteristics to information processing in the inequality of collaborating frequencies among the partners.

5. Conclusions

This study analyzed the change of each author's co-authorship relations according to increase in the number of papers, and described the characteristics of the co-authorship networks of the four domains. The differences among the domains, which are shown in this study, seem to be caused by the differences in research styles, basically. On the other hand, there is another possibility that the co-authorship networks are affected by undesirable constraints. For example, a difference in affiliation of authors may put obstacles in the way of their collaboration. In future works, we will take into consideration authors' affiliation, and analyze co-authorship networks more minutely.

References

Arunachalam, S., Srinivasan, R. & Raman, V. (1994). International collaboration in science: participation by the Asian giants. Scientometrics, 30(1), 7-22.

Chitashvili, R.J.,& Baayen, R.H. (1993). Word frequency distributions. In: Hrebicek,

L.,& Altmann, G. eds., Quantitative text analysis, 54-135. Trier: Wissenschaftlicher Verlag.

Devilliers, F. P. R. (1984). Publish or perish: the growing trend towards multiple authorship. South African Medical Journal, 66(23), 882-883.

Drenth, J. P. H. (1998). Multiple authorship: the contribution of senior authors. Journal of the American Medical Association, 280(3), 219-221.

Egghe, L. (2002). Sampling and concentration values of incomplete bibliographies. Journal of the American Society for Information Science and Technology, 53(4), 271-281.

Elalami, J., Dore, J. C.,& Miquel, J. F. (1992). International scientific collaboration in arab countries. Scientometrics, 23(1), 249-263.

Kageura, K. (1999). Some characteristics of bibliometric samples: an examination of Lotka-type data. Annals of Japan Society of Library Science, 44(3), 97-110.

Kretschmer, H.,& Gupta. B. M. (1998). Collaboration patterns in theoretical population genetics. Scientometrics, 43(3), 455-462.

Kundra, R.,& Kretschmer, H. (1999). A new model of scientific collaboration Part 2: collaboration patterns in Indian medicine. Scientometrics, 46(3), 519-528.

Narin, F., Stevens, K.,& Whitlow E. S. (1991). Scientific cooperation in Europe and the citation of multinationally authored papers. Scientometrics, 21(3), 313-323.

Okubo, Y., Miquel, J. F., Frigoletto, L.,& Dore, J. C. (1992). Structure of international collaboration in science: typology of countries through multivariate techniques using a link indicator. Scientometrics, 25(2), 321-351.

O'Neill, G. P. (1998). Authorship patterns in theory based versus research based journals. Scientometrics, 41(3), 291-298.

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. Journal of Information Science, 28(6), 441-453.

Tweedie, F.J.,& Baayen, R.H. (1998). How variable may a constant be?: measures of lexical richness in perspective. Computers and the Humanities, 32, 323-352.

Wasserman, S., & Faust, K. (1994). Social Network Analysis. New York, Cambridge University Press.

Yitzhaki, M., & Ben-Tamar, D. (1990). Multiple authorship in biochemistry and other fields: a case study of the Journal of Biological Chemistry throughout 1905-1988. In: Egghe, L., & Rousseau, R. eds., Informetrics, 89/90, 373-389. Amsterdam: Elsevier.

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), 18-32 [in Japanese].

	N	A(N)	C_L
Electrical Engineering	19784	25230	0.241
Information Processing	27047	24267	0.225
Polymer Science	21505	16820	0.213
Biochemistry	17782	21315	0.229

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

	$V_{av}(N)$	$G_{av}(N)$	C_{Lav}
Electrical Engineering	6.46	0.0755	0.256
Information Processing	4.75	0.0787	0.218
Polymer Science	6.21	0.109	0.226
Biochemistry	7.11	0.0958	0.257

Table 2. The co-authorship statistics for four domains

	C
Electrical Engineering	3.83
Information Processing	2.93
Polymer Science	3.55
Biochemistry	4.05

Table 3. The number of coauthors per paper in each domain

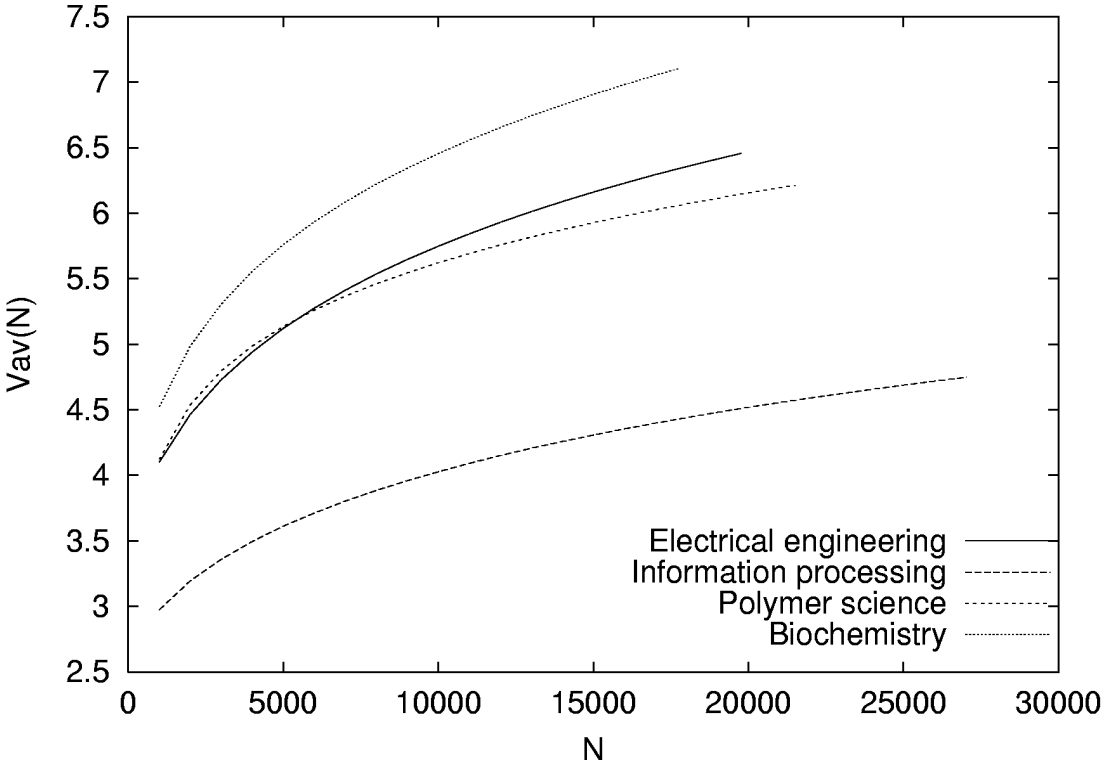

Figure 1a. The developmental profile of $V_{av}(N)$ in each domain

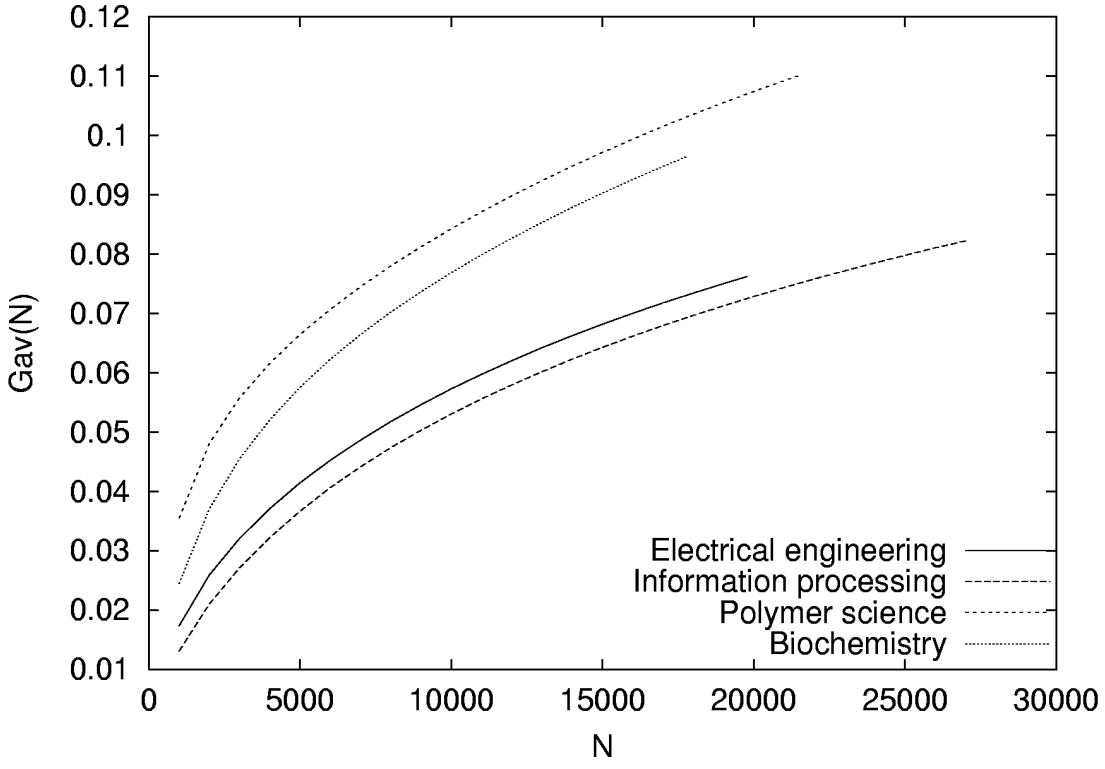

Figure 1b. The developmental profile of $G_{av}(N)$ in each domain

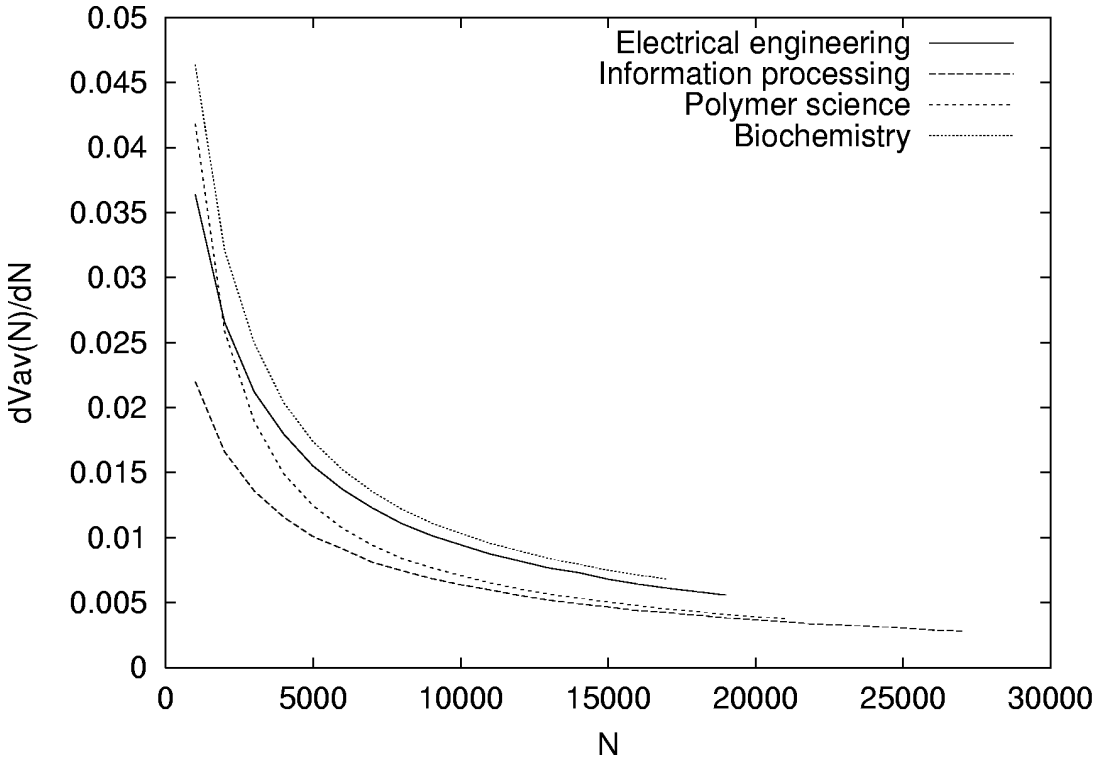

Figure 2a. The growth rate of $V_{av}(N)$ in each domain

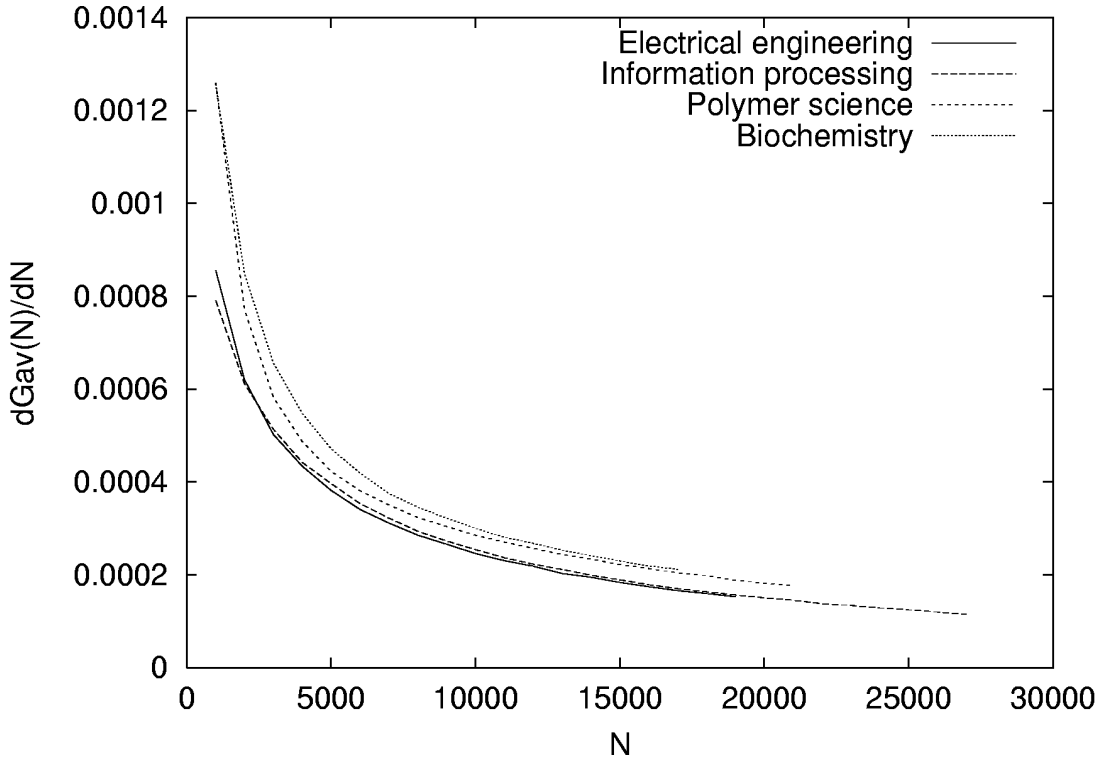

Figure 2b. The growth rate of $G_{av}(N)$ in each domain