Comparative Analysis of Co-authorship Networks Considering Authors' Roles in Collaboration: Differences between the Theoretical and Application Areas

Fuyuki Yoshikane*, Takayuki Nozawa* and Keita Tsuji**

*fuyuki@niad.ac.jp, nozawa@niad.ac.jp
Faculty of University Evaluation and Research, National Institution for Academic Degrees and University Evaluation, 1-29-1 Gakuen-nishimachi, Kodaira, Tokyo 187-8587 (Japan)

**keita@nii.ac.jp
National Institute for Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 (Japan)

Abstract
Many studies have analyzed "direct" partnerships in co-authorship networks. One the other hand, the whole network structure, including "indirect" links between researchers, has not been sufficiently studied yet. This study aims at deriving knowledge about the communication structures regarding production of papers by analyzing the researchers' activities from different viewpoints considering roles in co-authorship networks. In this study, we compare the co-authorship networks between the theoretical and application areas in computer science. By applying the modified HITS algorithm to the co-authorship networks, we analyze for each researcher in the co-authorship networks (1) the degree of importance as the leader and (2) that as the follower. We further examine the correlation between these two viewpoints. This study has shown that the negative correlation between (1) and (2) is greater in the application area. It suggests that, in computer science, the two roles (i.e., the leader and the follower) are more clearly separated from each other in the application area than in the theoretical area.

Introduction

In academic research, it is exceedingly rare that a researcher produces outcomes with no connection to the context of the research community. New findings are usually derived from the context of the research community, that is, from the accumulation of preceding researches or cooperative relationships in the research domain. Therefore, when we analyze the activity of the researchers in some domain for the purpose of grasping the characteristics of that domain in producing knowledge, we must not only evaluate each researcher's activity individually, but also take into consideration his/her position in the structure of some kind of intellectual tie.

In recent years, certain factors, such as the specialization of researchers and the growth of interdisciplinary fields, have caused researchers to collaborate (Sacco & Milana, 1984; Muñoz & Moore, 1985; Cason, 1992; Andersson & Persson, 1993; Bird, 1997; Bordons & Gómez, 2000). Now, collaborating with colleagues (i.e., synchronic networking), as well as citing preceding researchers' outcomes (i.e., diachronic networking), is very important in research activities. There are a large number of studies which deal with collaboration (co-authorship) networks. Most of those chiefly analyzed "direct" partnerships in collaboration networks. For example, many studies examined how researchers' attributions, such as productivity, status and gender, influence their preferences in choosing collaboration partners (e.g., Kretschmer, 1994; 1997; Bahr & Zemon, 2000; Yoshikane & Kageura, 2004). One the other hand, some studies examined the whole structure of the collaboration network (e.g., Kretschmer, 2004). However, the whole network structure, including "indirect" links between researchers who are not each other's partners but have common partners, has not been sufficiently studied yet. By analyzing the researchers' activities from different viewpoints considering roles in collaboration networks, this study aims at deriving knowledge about the communication structures of research communities.

The idea that the researchers' activities should be understood in the context of the research community is also regarded as important in research evaluation. For instance, the National Institution for Academic Degrees and University Evaluation, Japan (NIAD-UE) adopts research collaboration and cooperation as an important viewpoint in evaluating research activities (NIAD-UE, 2003). Research activities must be evaluated from various viewpoints, including the roles in collaboration networks,

Proceedings of ISSI 2005, pages 509–516.
Edited by P. Ingwersen and B. Larsen.

and different viewpoints may yield different results in research evaluation. In this study, we examine the correlation between measures corresponding to some viewpoints in research evaluation, and show that those measures are not necessarily positively correlated.

For this purpose, this study analyzes co-authorship networks of two different domains. Cooperative relationships in research activities are not observed only in authorship credits of coauthored papers. Some of them are observed in acknowledgments of papers and not in co-authorship credits. However, we assume that co-authorship credits cover all collaborators that "substantially and technically" contribute to their coauthored papers while acknowledgments are addressed only to subsidiary supporters[1]. So this study measures the activity of research collaboration by analyzing co-authorship networks observed in published papers.

There is another problem concerning co-authorship, that is, honorific authorship, by which co-authorship credits are sometimes regarded to be irresponsible (Cason, 1992). Although many studies have pointed out the problem of honorary coauthors that have no substantial contribution to the work, ethical guidelines regarding authorship issues have been laid down in recent years in each domain (e.g., ICMJE, 1997). Those guidelines state that authorship credits should be determined by substantial and technical contributions to the work. Some questionnaire surveys illustrate that the majority of researchers reached a consensus following those guidelines (Hoen, Walvoort & Overbeke, 1998; Bartle, Fink & Hayes, 2000). Taking into account this situation, we reasonably assume that co-authorship credits represent the substantial and technical contributions.

The target domains whose networks are compared in this study are two subdomains in computer science. There are two reasons why we chose computer science. One is that researchers in this domain perform research collaboration very actively. Hence, the analysis of their collaboration networks is deemed to be very important. The other reason is that, as not only the theoretical research area but also the interdisciplinary application area is flourishing in this domain, it is expected that we will obtain useful knowledge about the correlation between the collaboration tendency and research style from the differences between the theoretical and application areas.

This paper is organized as follows. First we explain the source data used in the analysis, and then, after narrating our viewpoints and methodology, we display the results of our experiments comparing the co-authorship networks of the two subdomains in computer science. Lastly, based on these results, we sum up the characteristics of each subdomain.

Data

The data used for observing co-authorship networks were extracted from *SCI* (*Science Citation Index*) CD-ROM version, provided by Thomson ISI. From the database, we extracted the records of papers published between 1999 and 2003. *SCI* is one of the most comprehensive bibliographic databases in natural sciences, though the journals contained in it are, for the most part, English-language ones. Besides, *SCI* selects only core journals that satisfy the qualitative criteria[2]. We therefore chose *SCI* as the source data in this study. *SCI* covers not only original papers but also various types of documents such as reviews, letters, and so on. As mentioned in the previous section, our interest is in the network structures of substantial and technical collaboration. Thus, as the target of observation, we extracted only original papers[3], which are considered to most directly reflect the structures of substantial and technical collaboration.

[1] For instance, Cronin, Shaw & Barre (2003) analyzed the acknowledgments of papers in psychology and philosophy, and showed that the acknowledgments are used for signifying subsidiary support rather than substantial and technical collaboration.

[2] On the basis of peer review and citation analysis, the quality of researches is evaluated. Not only the quality of researches but also the international and geographic diversity among authors of papers included in the journal is taken into consideration (Testa, 2004).

[3] We extracted records whose "document type" fields are "article". *SCI* includes "meeting-abstract", "letter", "review", "software-review", "biographical-item", "editorial-material", etc., besides "article".

This study adopted the category classification of "List of source publications: arranged by subject category" in *SCI*. According to it, we selected the core journals to be analyzed for each of the two target domains, "the theoretical area in computer science" and "the application area in computer science". Henceforth for succinctness, we call them "the theoretical area" and "the application area", respectively. From the database, we extracted the bibliographic data of the papers of the journals included in the two categories, "computer science, theory & methods" and "computer science, interdisciplinary applications", as the data of the theoretical and application areas. 21 journals (e.g., *Journal of Algorithms*) were extracted for the theoretical area, and 22 journals (e.g., *Computer Applications in the Biosciences*) were extracted for the application area. In this study, on the basis of the category of the journal where the researcher's papers appear, the researcher is connected to the domain. That is to say, all authors whose papers appear in the journals classified to "computer science, theory & methods (or interdisciplinary applications)" in *SCI* are regarded to belong to the theoretical area (or the application area).

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

	NJ	*NP*	*TA*	*DA*	A_{av}	P_{av}
Theoretical area	21	9663	22485	14525	2.33	1.55
Application area	22	11584	32341	21801	2.79	1.48

Table 1 shows the basic quantities of data in each of the theoretical and application areas. There is not much of a difference between them in the number of journals *NJ*, the number of papers *NP*, the average number of authors per paper A_{av} (= *TA*/*NP*), and the average number of papers per author P_{av} (= *TA*/*DA*). With regard to the number of authors, the application area is about one and half times larger than the theoretical area, both in the total number of author tokens *TA* and in the number of different authors *DA*. As far as judging from the data, we can state that, in computer science, the application area consists of more researchers than the theoretical area.

Methodology

Viewpoints

This study analyzes the researchers' activity of producing papers from two viewpoints, (1) the degree of importance as the first author and (2) that as the coauthor, excluding the first author, in co-authorship networks. We differentiate between these two viewpoints, on the basis of the assumption that the first author designs the whole research as the leader and plays the special role, which is different from other coauthors' roles. In some domains, such criteria are specified in guidelines, and we also see this in the results of awareness surveys (e.g., Bridgwater, Bornstein & Walkenbach, 1981). Thus, we can regard this assumption to be reasonable, at least to some extent. This study analyzes the two viewpoints, (1) and (2), on the basis of the idea that both the roles, as a leader in producing papers and as a follower collaborating with the leader, are important in the network of research collaboration, and that these two roles are essentially different.

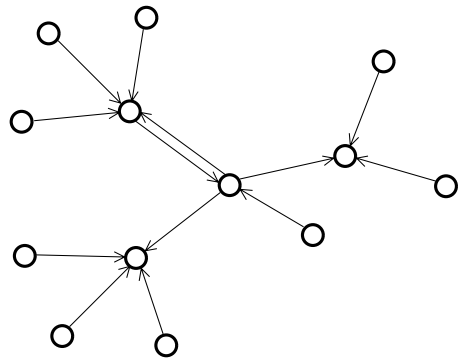

Figure 1: An example of a co-authorship network

We assume the following model, for setting the operational definitions and measures of the above two viewpoints.

- Assuming directed graphs where the ties (links) are oriented from coauthors to the first author for each paper (see Figure 1).
- Assuming weighted graphs where the strength of co-authorship relations is taken into account.

Many indices have been proposed for measuring the strength of co-authorship relations (e.g., Narin, Stevens & Whitlow, 1991; Arunachalam, Srinivasan & Raman, 1994). This study uses the coauthoring frequency itself as the weight of ties in graphs simply, on the basis of the assumption that the strength of co-authorship relations between a pair of researchers grows in proportion to the number of times they have published co-authored papers. By applying the HITS algorithm, which will be introduced in the next section, to the above-mentioned weighted directed graphs, we calculate the degree of importance for each node (researcher) in consideration of the global structures of networks.

The modified HITS algorithm

In this study, we calculate the degree of importance of each researcher in the network of research collaboration, giving attention to the number of collaborating partners, the relationship strength with each partner, and moreover the degree of importance of each partner. The degree of importance as the leader $C_l(n_i)$ and that as the follower $C_f(n_i)$ are obtained for each researcher n_i by the following formulas, respectively.

$$C_l(n_i) = \sum_{j=1}^{g} a_{ji} C_f(n_j) \qquad (1)$$

$$C_f(n_i) = \sum_{j=1}^{g} a_{ij} C_l(n_j) \qquad (2)$$

where g represents the number of nodes in the network, that is, the number of researchers in the domain. a_{ij} represents the value in cell (i, j) of the adjacency matrix A of the co-authorship network, and is equal to the relationship strength of the tie oriented from n_j to n_i, that is, the number of coauthored papers where n_i is the first author and n_j is his coauthor. (The value of diagonal cells a_{ii} is 0.)

Here, we assume the mutual dependency that "a researcher who assists important leaders plays an important role as the follower, and a researcher who organizes important followers plays an important

role as the leader". In the formulas (1) and (2), by repeating recursive substitution, the global structure of the co-authorship network is reflected in the degree of importance of each researcher. This recursive substitution results in solving the eigenvector problem of the adjacency matrix A.

The common idea that ties with more important nodes contribute to the degree of importance more than those with less important ones is shared among the centrality measure of Bonacich (1987), the PageRank algorithm (Brin & Page, 1998), and the HITS (Hyperlink-Induced Topic Search) algorithm (Kleinberg, 1998)[4]. Among the three, the HITS algorithm is most similar to the measures used in this study, in that both of these assume two different roles considering the direction of relationships. In this study, in addition to the direction of relationships, the strength (weight) of relationships is reflected in the calculation of the measures. The co-authorship networks to be analyzed in this study have so many nodes (researchers) that it is hardly able to solve the eigenvector equations of their adjacency matrices. Thus, we calculate $C_l(n_i)$ and $C_f(n_i)$ by recursively repeating substitution and normalization of vectors in the same manner as the HITS algorithm[5].

Results

For each of the application and theoretical areas in computer science, the degree of importance as the leader $C_l(n_i)$ and the degree of importance as the follower $C_f(n_i)$ are calculated for each researcher. The correlation between these two types of importance measures is shown in Table 2. In this study, we used the Spearman's rank correlation coefficient r_s, because both of the two measures do not follow the normal distribution and these are not expected to be linearly associated with each other. In addition to the correlation coefficient, Table 2 shows the average values of the degree (outdegree or indegree) per node D_{av} in those co-authorship networks.

Table 2. The co-authorship network properties for the two domains.

	D_{av}	r_s
Theoretical area	1.60	-0.438
Application area	1.72	-0.488

The average value of the degree per node represents how many leaders or followers a researcher is linked with on average. In the average value of the degree per node D_{av}, as well as in the average number of authors per paper A_{av} shown in Table 1, the application area is slightly higher than the theoretical area. It suggests that researchers in the application area have more cooperative relationships than do those in the theoretical area. A negative correlation between the two importance measures, $C_l(n_i)$ and $C_f(n_i)$, is observed, both in the theoretical and application areas. That is to say, there is a tendency that the roles as the leader and as the follower are played by different researchers rather than that the same researcher plays both of these two roles. As the negative correlation is greater in the application area, it is suggested that the two roles are separated from each other in the application area more clearly than in the theoretical area. A possible reason for this is that, in application area, there are more "peripheral" researchers whose main fields are the targets of application (chemistry, medicine, geoscience and so on) rather than computer science itself. They might collaborate only as followers.

Tables 3 and 4 show the characteristics of researchers ranked as the most important leaders and the most important followers. D_{in} and D_{out} represent the indegree and the outdegree of each node (researcher). While P represents the number of published papers of each researcher, P_{sin}, P_1, P_{2-}, and P_{last} represent the number of his single-authored papers, the number of multiple-authored papers where he is the first author, the number of multiple-authored papers where he is not the first author but a coauthor, and the number of multiple-authored papers where his name is listed last, respectively. There is no researcher that is listed in the top ten as both the leader and the follower, except two researchers (a researcher in the theoretical area who is ranked as the third most important leader and

[4] The HITS and PageRank algorithms were devised for the scoring of results of web page searches.
[5] Substitution and normalization are repeated 10 times.

the second most important follower, and a researcher in the application area who is ranked as the tenth most important leader and the ninth most important follower).

The modified HITS algorithm used in this study calculates scores on the basis of the position in collaboration networks. By this algorithm, we aim to measure not researchers' productivity but the degree of importance in collaboration networks. So, it assigns high scores to researchers who occupy important positions with linkages to important researchers, whether they themselves publish many papers or not[6].

Table 3. The characteristics of researchers ranked as the most important leaders.

	Theoretical area							Application area						
Rank	D_{in}	D_{out}	P	P_{sin}	P_1	P_{2-}	P_{last}	D_{in}	D_{out}	P	P_{sin}	P_1	P_{2-}	P_{last}
1	41	1	20	1	18	1	1	69	0	2	0	2	0	0
2	6	2	12	1	8	3	1	20	2	14	0	12	2	1
3	18	1	13	0	9	4	0	5	2	16	1	9	6	0
4	1	0	3	1	2	0	0	6	2	18	0	8	10	0
5	5	5	7	0	2	5	1	2	0	1	0	1	0	0
6	3	0	2	0	2	0	0	12	8	33	0	23	10	9
7	21	0	9	0	9	0	0	3	2	10	0	4	6	0
8	8	1	10	0	9	1	0	3	0	2	0	2	0	0
9	2	1	2	0	1	1	0	3	0	1	0	1	0	0
10	2	2	3	0	1	2	1	3	1	3	0	1	2	0

A major difference between the theoretical and application areas is observed in the characteristics of followers (see Table 4). In the theoretical area, there are researchers who not only often collaborate as the follower with a few specific leaders (i.e., D_{out} is not zero but very small, and P_{2-} is large), but also publish coauthored papers as the leader actively (i.e., P_1 is large). The second most important follower mentioned above is a typical example of this type. On the other hand, there is another type of important follower in the theoretical area. This type of follower collaborates with various leaders (i.e., D_{out} is large) as the supervisor (i.e., P_{last}/P is large)[7]. The first, sixth and ninth most important followers in the theoretical area are typical examples of this type.

The latter type is assumed to play a role as a kind of coordinator who arranges research groups, bringing over proper specialists for research projects on the basis of his own connections. It seems that, while leaders function as "hubs" in the networks of knowledge communication, those coordinators function as "bridges" which intermediate between the "hubs" (e.g., the node (a) in Figure 2). In the application area, by contrast, this kind of coordinator does not appear as an important follower in Table 4. The application area in computer science is close to "Mode 2" (Gibbons et al., 1994) in which researches are transdisciplinary and are carried out in the context of application arising from society rather than within the discipline. Therefore, it may be assumed that persons who play the role as the "coordinator" in the application area often exist outside the domain and have no substantial and technical cooperative relationship with "leaders", and that it is reflected in the co-authorship network in this domain.

[6] However, some extreme instances are observed in Table 3. That is, two researchers with only one paper are ranked as the most important leaders. It might be necessary for more reasonable scoring scheme to refine the method of weighting.

[7] In many cases, the name of the supervisor is listed last in coauthored papers. For example, in the case of students' works, their faculty advisors are often listed last as the supervisor.

Table 4. The characteristics of researchers ranked as the most important followers.

	Theoretical area							Application area						
Rank	D_{in}	D_{out}	P	P_{sin}	P_1	P_{2-}	P_{last}	D_{in}	D_{out}	P	P_{sin}	P_1	P_{2-}	P_{last}
1	2	9	18	0	1	17	16	0	3	4	0	0	4	2
2	18	1	13	0	9	4	0	0	2	3	0	0	3	0
3	0	6	8	0	0	8	0	0	2	3	0	0	3	1
4	1	4	8	0	1	7	2	1	2	4	0	1	3	0
5	1	3	9	3	1	5	2	0	2	3	0	0	3	0
6	0	7	10	0	0	10	6	0	2	3	0	0	3	0
7	0	1	8	0	0	8	2	0	2	3	0	0	3	1
8	0	3	4	0	0	4	0	0	2	3	0	0	3	1
9	0	5	7	0	0	7	6	3	1	3	0	1	2	0
10	1	3	9	4	1	4	1	1	1	3	0	1	2	0

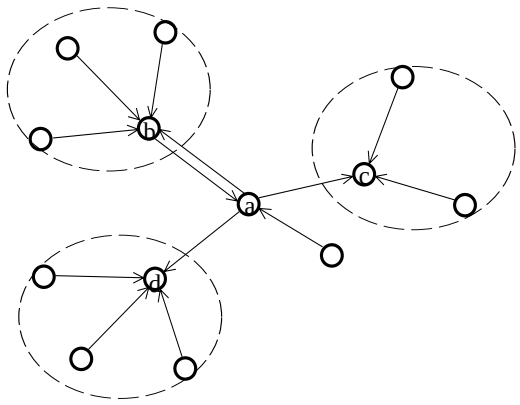

Figure 2: A coordinator and leaders

Conclusions

This study has compared the co-authorship networks of the two subdomains in computer science, that is, the theoretical and application areas, from two viewpoints giving attention to the roles in collaboration networks. By applying the modified HITS algorithm to the co-authorship networks, we analyzed (1) the degree of importance as the leader and (2) that as the follower, for each researcher in the co-authorship networks. Subsequently, the correlation between these two viewpoints was examined.

It was shown that the negative correlation between (1) and (2) is greater in the application area. This result suggests that, as for computer science, the two roles (i.e., the "leader" and the "follower") are separated from each other in the application area more clearly than in the theoretical area. Moreover, we indicated the differences between the two subdomains regarding the characteristics of researchers occupying the most important positions in the co-authorship networks. Strictly speaking, of course, the differences among the theoretical and application areas shown in this study can be regarded just as the features in computer science. In future studies, we will analyze co-authorship networks of other domains, and examine whether these differences can be generalized.

References

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

Andersson, A. E. & Persson, O. (1993). Networking scientists. *Annals of Regional Science*, 27(1), p. 11-21.

Bahr, A. H. & Zemon, M. (2000). Collaborative authorship in the journal literature: Perspectives for academic librarians who wish to publish. *College & Research Libraries*, 61(5), 410-419.

Bartle, S. A., Fink, A. A. & Hayes, B. C. (2000). Psychology of the scientist: LXXX. attitudes regarding authorship issues in psychological publications. *Psychological Reports*, 86(3), part 1, 771-788.

Bird, J. E. (1997). Authorship patterns in marine mammal science, 1985-1993. *Scientometrics*, 39(1), 99-105.

Bonacich, P. (1987). Power and centrality: A family of measures. *The American Journal of Sociology*, 92(5), 1170-1182.

Bordons, M. & Gómez, I. (2000). Collaboration networks in science. In B. Cronin & H. B. Atkins (Eds.), *Web of Knowledge: A Festschrift in Honor of Eugene Garfield* (pp. 197-213). Medford: Information Today.

Bridgwater, C. A., Bornstein, P. H. & Walkenbach, J. (1981) Ethical issues and the assignment of publication credit. *American Psychologist*, 36(5), 524-525.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of 7th International World Wide Web Conference* (pp. 101-117). Amsterdam: Elsevier Science.

Cason, J. A. (1992) Authorship trends in poultry science, 1981 through 1990. *Poultry Science*, 71(8), 1283-1291.

Cronin, B., Shaw, D. & Barre, K. L. (2003). A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *Journal of the American Society for Information Science and Technology*, 54(9), 855-871.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P. & Trow, M. (1994). *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. London: Sage Pub.

Hoen, W. P., Walvoort, H. C. & Overbeke, A. J. P. M. (1998). What are the factors determining authorship and the order of the authors' names?: A study among authors of the Nederlands Tijdschrift voor Geneeskunde (Dutch Journal of Medicine). *Journal of the American Medical Association*, 280(3), 217-218.

ICMJE (International Committee of Medical Journal Editors) (1997). Uniform requirements for manuscripts submitted to biomedical journals. *Journal of the American Medical Association*, 277(11), 927-934.

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of 9th Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 668-677). New York: Association for Computing Machinery.

Kretschmer, H. (1994). Coauthorship networks of invisible-colleges and institutionalized communities. *Scientometrics*, 30(1), 363-369.

Kretschmer, H. (1997). Patterns of behaviour in coauthorship networks of invisible colleges. *Scientometrics*, 40(3), 579-591.

Kretschmer, H. (2004). Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the Web. *Scientometrics*, 60(3), 409-420.

Muñoz, W. P. & Moore, P. J. (1985). Multiple authorship. *South African Medical Journal*, 68(6), 368.

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

NIAD-UE (National Institution for Academic Degrees and University Evaluation, Japan) (Ed.). (2003). *The Scheme for Implementation of Self-Evaluation: University-Wide Thematic Evaluation, International Cooperation and Exchange Activities (University Evaluations Begun in Fiscal Year 2002)*, rev. ed. Tokyo: NIAD-UE [in Japanese].

Sacco, W. P. & Milana, S. (1984). Increase in number of authors per article in 10 APA journals: 1960-1980. *Cognitive Therapy and Research*, 8(1), 77-83.

Testa, J. (2004). The ISI database: The journal selection process. *ISI Essay*. Retrieved November 2, 2004 from: http://www.isinet.com/essays/selectionofmaterialforcoverage/199701.html.

Yoshikane, F. & Kageura, K. (2004). Comparative analysis of coauthorship networks of different domains: The growth and change of networks. *Scientometrics*, 60(3), 433-444.