Structural Analysis of Paper Citation and Co-Authorship Networks using Network Analysis Techniques

Kouhei Sugiyama, Hiroyuki Ohsaki and Makoto Imase
Graduate School of Information Science and Technology,
Osaka University, Japan
E-mail: {k-sugi,oosaki,imase}@ist.osaka-u.ac.jp

1 Introduction

In recent years, attention to structure of real complex networks has been increasing [1]. For instance, researches on social networks, which represents communications among people, has been actively performed [3–5]. Using several network-analysis techniques, structure of real several networks, such as hyperlink structure of Web pages and co-authorship relation of research, has been actively analyzed [1, 2, 14]. For instance, in [14], a movie actors network (network representing costarring relation in movies) is analyzed. Consequently, it is shown that in the movie actors network, the network distance (maximum of shortest path lengths) between arbitrary actors (i.e., node) is small, and that the movie actors network has clustered structure. Moreover, in [2], the mathematics co-authorship network is analyzed. Consequently, it is shown that in the mathematics co-authorship network, the number of co-authors is small (approximately three), and that the mathematics co-authorship network has clustered structure.

In addition, researches on a mechanism determining structure of real networks and cause of small-world and scale-free structures in real networks have been performed [7, 8]. For instance, in [7], a simple network generation model for reproducing characteristics of real social networks is proposed.

Such characteristics of network structure could be utilized in various ways. For instance, it is expected that an advanced information retrieval system, rather than conventional keyword-based information retrieval systems, can be realized by utilizing characteristics of the network structure. In recent years, several information retrieval systems using network structure are proposed [11]. In the literature, however, insufficient investigation on what the network structure tells us and how the network structure can be utilized has been performed.

In this paper, we focus on a paper citation network and a paper co-authorship network as one of real complex networks, and investigate what we can know from the structure of these complex networks. The *paper citation network* is a network expressing citation relation among scientific papers. The *paper co-authorship network* is a network expressing existence of co-authorship relation between authors of scientific papers. By utilizing network-analysis techniques, we investigate the structure of paper citation/co-authorship networks.

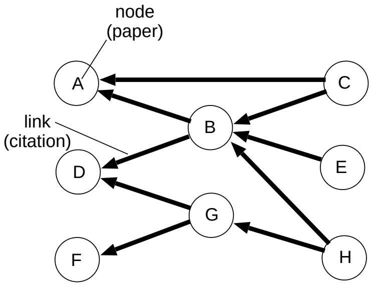

Figure 1: Example of a paper citation network

In this paper, we first build paper citation/co-authorship networks by analyzing citation relation and co-authorship relation of a large number of scientific papers. Specifically, we extract citation relation and co-authorship relation from scientific papers database [6], and build paper citation/co-authorship networks. We then analyze the structure of paper citation/co-authorship networks by applying several network-analysis techniques. Consequently, we show that the paper citation network has less grouped structure than other social networks, and that the distribution of node in-degrees (i.e., the number of papers cited) follows a power-law. We also show that the paper co-authorship network has grouped structure and exhibits scale-free structure.

The organization of this paper is as follows. Section 2 explains how the paper citation network was built from the scientific papers database. By using several network-analysis techniques, clustering coefficient, and degree distribution of the paper citation network are also calculated. In Section 3, we build the paper co-authorship network using a similar approach to Section 2, and investigate its characteristics using several network-analysis techniques. Finally, Section 4 concludes this paper and discusses future works.

2 Paper Citation Network

2.1 Network Construction

In this paper, a paper citation network is defined as a network expressing citation relation among scientific papers. Generally, a paper cites many papers, and those cited papers are listed as the bibliography at the end of a paper. In a paper citation network, a paper corresponds to a node and citation relation between two papers corresponds to a link.

An example of a paper citation network is shown in Fig. 1. As shown in Fig. 1, paper A and paper B are represented by node A and node B. If paper B cites paper A, this citation relation is represented by a directed link from node B to node A. By representing citation relations as a network, it becomes possible to analyze the structure of citation relations.

From the bibliographic information of scientific papers from the scientific papers database [6], we built the paper citation network. This scientific papers database includes information on thousands of journal papers and international conference papers published by the IEEE Communications Society during 1952–2004. This database contains various information, such as title, authors, keywords, and bibliography of all papers.

A paper included in the scientific papers database is assigned as a node in the paper citation network. From the bibliography in the database, directed links are generated from the paper to all cited papers in the paper citation network. Note that the paper citation network is a directed graph. The bibliography sometimes contains entries of papers not included in the scientific papers database. The paper citation network does not contain links corresponding to those external papers.

Table 1: Characteristics of paper citation/co-authorship networks and other social networks

	N	$\overline{k}$	C
paper citation network	27,865	3.69	0.24
paper co-authorship network	24,841	4.76	0.76
hyperlink structure of Web pages [10]	269,504	5.55	0.29
AS-level topology in the Internet [10]	10,697	5.98	0.39
mathematics co-authorship network [2]	70,975	3.9	0.59
physics co-authorship network [9]	52,909	9.27	0.56
movie actors network [14]	225,226	61	0.79

The fundamental characteristics of the paper citation network is summarized in Tab. 1. This table shows that the number of nodes N is 27,865, and the average degrees $\overline{k}$ is 3.69. The clustering coefficient C are also included in the table. Definitions of those values will be explained in the following Section.

2.2 Clustering Coefficient

We investigate the clustering coefficient of the paper citation network. We investigate clustered structure of the paper citation network. The objective of this section is to answer the following questions: does each paper belong to a specific group (e.g., a specific research field)? If so, are the citation relations closed within those groups?

In Section 2.1, the paper citation network was expressed as a directed graph. In what follows, the paper citation network is expressed as a undirected graph to investigate clustering coefficient.

Clustering coefficient is a metric that measures cluster structure of nodes in a network. Specifically, clustering coefficient is defined as a probability that triangular cycle exists between each node pair. The clustering coefficient C of a network is defined as

$$C \equiv \frac{1}{N} \sum_{i=1}^{N} C_i, \tag{1}$$

where C_i is the clustering coefficient of node i, which is defined as

$$C_i \equiv \frac{2E_i}{k_i(k_i - 1)}. \tag{2}$$

In the above equation, k_i is the degree of node i, and E_i is the total number of triangular cycles, which start from and return to node i. Let N be the number of nodes, and $\overline{k}$ be the average degree. It is known that the clustering coefficient C_{rand} of a random network is given by [1]

$$C_{rand} = \frac{\overline{k}}{N}. \tag{3}$$

Figure 2 shows the relation between the number of nodes and the clustering coefficient (i.e., clustering coefficient C normalized by degree $\overline{k}$) of the paper citation network. In the figure, the clustering coefficient of a random network, hyperlink structure of Web pages (web hyperlink) [10], and AS-level topology in the Internet (Internet) [10] are also shown. This figure shows that the clustering coefficient C of the paper citation network is larger than the clustering coefficient C_{rand} of a random network. This means that nodes (i.e., papers) tend to have more grouped structure in the paper citation

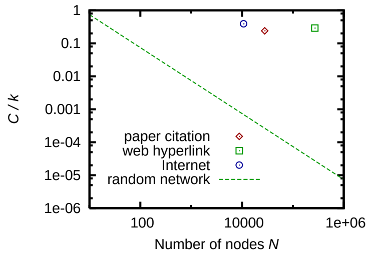

Figure 2: Clustering coefficient of the paper citation network (dotted line is the case of random network)

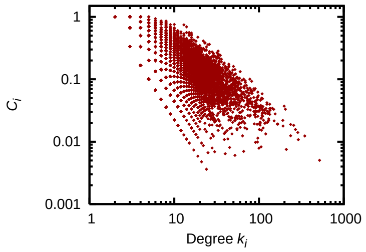

Figure 3: Paper citation network (degree vs. local clustering coefficient)

network than in a random network. However, if compared with the hyperlink structure of Web pages, and the AS-level topology in the Internet, the value of clustering coefficient is small.

As can be seen in Eq. (1), clustering coefficient C is defined as the average value of clustering coefficients of all nodes (local clustering coefficient). Figure 3 shows the relation between the degree of each node and its local clustering coefficient. This shows that local clustering coefficient C_i tends to become small as the degree k_i of each node becomes large. This result suggests that a node with a large degree (i.e., a popular paper) is likely to be cited by many papers.

2.3 Degree Distribution

We then investigate degree distribution of the paper citation network. Namely, we investigate the distribution of the number of paper citation and paper cited of each node (i.e., paper) in the paper citation network. The objective of this section is to answer the following questions: does any hub paper, which has a very large number of paper citations with other papers, exist in the paper citation network?

The degree distribution is determined by the degree-distribution function $P(k)$. $P(k)$ is a probability that the degree of a randomly chosen node is k in a network.

When the number of nodes N is large, it is known that the degree-distribution function $P(k)$ of a random network with the average degree $\overline{k}$ approximately follows the Poisson distribution [1]; i.e.,

$$P(k) \simeq e^{-\overline{k}} \frac{\overline{k}^k}{k!}. \tag{4}$$

It is reported that many real social networks have a scale-free property unlike a random network [12]; i.e., the tail of the degree-distribution function $P(k)$ is governed by a

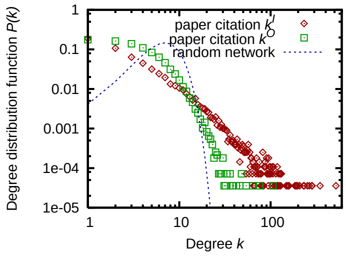

Figure 4: Degree distribution of the paper citation network (dotted line is the case of random network)

power-law function

$$P(k) \sim k^{-\gamma}, \tag{5}$$

where γ is a constant and called a *power-law index*. It is known that many real scale-free networks have a power-law index γ of 2.0–3.4 [1].

Figure 4 shows the degree distribution of the paper citation network. Figure 4 shows the estimated degree distribution function $P(k)$ as a function of node in-degree k^I and out-degree k^O. The dotted line in the figure shows the degree distribution function of a random network.

It can be found from this figure that the in-degree distribution of the paper citation network is considerably different from the that of a random network. Also can be found that the tail of the degree distribution function $P(k)$ follows a power law. The average degree $\overline{k}$ is comparatively small (i.e., 3.69). We also found that a few hub nodes whose in-degrees are 500 or more exist. We found that the power-law index γ of the paper citation network was $\gamma = 1.8$. It can be found that the in-degree distribution of the paper citation network is considerably different from that of a random network. However, it can be found that the tail of the degree distribution function $P(k)$ does not follows a power law.

3 Paper Co-Authorship Network

3.1 Network Construction

The paper co-authorship network is a network expressing existence of co-authorship relation between authors of scientific papers. Co-authorship relations are relations representing whether an author have written a paper with another author in the past. Typically, a paper is written by two or more authors. In the paper co-authorship network, each author corresponds to a node and co-authorship relation among papers corresponds to a link.

An example of a paper co-authorship network is shown in Fig. 5. As shown in Fig. 5, author A and author B are represented by node A and node B. When a paper is written by author A, author B, and author C, these co-authorship relations among authors are represented by undirected links between node A and node B, between node A and node C, and between Node B and Node C. By representing co-authorship relations as a network, it becomes possible to analyze the structure of co-authorship relations.

In this paper, we obtained the information on co-authorship relation of papers from the scientific papers database [6] explained in Section 2.1, and built the paper co-authorship network. Refer to Section 2.1 for the details of this database.

Each author included in the scientific papers database is assigned as a node in the paper co-authorship network. Based on the author list in the database, undirected links

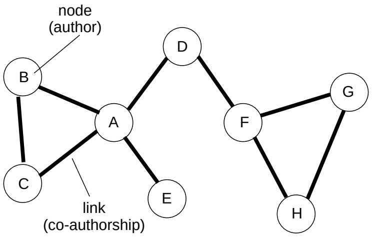

Figure 5: Example of a paper co-authorship network

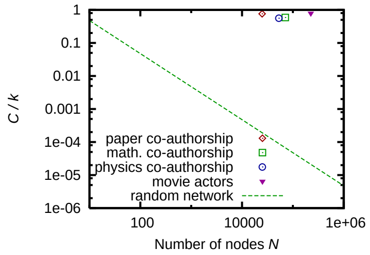

Figure 6: Clustering coefficient of the paper co-authorship network (dotted line is the case of random network)

between authors in the paper co-authorship network are generated. We built the paper co-authorship network as an undirected graph.

The fundamental characteristic of the paper co-authorship network is shown in Tab. 1. This table shows that the number of nodes N is 24,841, and the average degree $\overline{k}$ is 4.76.

3.2 Clustering Coefficient

We investigate the clustering coefficient of the paper co-authorship network. We investigate clustered structure of the paper co-authorship network. The objective of this section is to answer the following questions: does each author belong to a specific research group? If so, are the co-authorship relations closed within those groups?

Figure 6 shows the relation between the number of nodes and the clustering coefficient (i.e., clustering coefficient C normalized by degree $\overline{k}$) of the paper co-authorship network. In the figure, the clustering coefficient of a random network, mathematics co-authorship-relation network (math. co-authorship) [2], physics co-authorship-relation network (physics co-authorship) [9], movie actors network (movie actors) [14]. are also shown. This figure shows that the clustering coefficient C of the paper co-authorship network is larger than the clustering coefficient C_{rand} of a random network. This means that nodes (authors) tend to have more grouped structure in the paper co-authorship network than in a random network. Moreover, if compared with the mathematics and physics co-authorship networks, the clustering coefficient of the paper co-authorship network is closed to the clustering coefficients of the mathematics and physics co-authorship networks.

As can be seen in Eq. (1), clustering coefficient C is defined as the average value of the clustering coefficients of all nodes (local clustering coefficient). Figure 7 shows the relation between the degree of each node and its local clustering coefficient.

This figure shows that there exist node with a large degree (e.g., more than 10) whose clustering coefficient C_i is 1.0. This means that several authors have co-

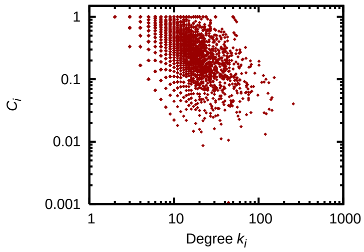

Figure 7: Paper co-authorship network (degree vs. local clustering coefficient)

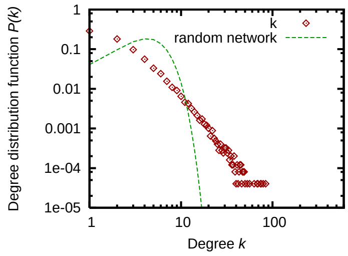

Figure 8: Degree distribution of the paper co-authorship network (dotted line is the case of random network)

authorship relation with more than 10 authors, and each of which has co-authorship relation each other, including existence of certain research groups.

3.3 Degree Distribution

We finally focus on degree distribution of the paper co-authorship network. We investigate the distribution of the number of co-authorship relation of each node (i.e., author) in the paper co-authorship network. The objective of this section is to answer the following questions: does any hub author, which has a very large number of co-authorship relations with other authors, exist in the paper co-authorship network? does the paper co-authorship network have a scale-free property?

Figure 8 shows the degree distribution of the paper co-authorship network. Figure 8 shows the estimated degree distribution function $P(k)$ as a function of node degree k.

It can be found from this figure that the degree distribution of the paper co-authorship network is considerably different from that of a random network. Also can be found that the tail of the degree distribution function $P(k)$ follows a power law. The average degree $\overline{k}$ is comparatively as small (i.e., 4.76). We also found that a few hub nodes whose degrees are more than 50 exist. We found that the power-law index γ of the paper co-authorship network is $\gamma = 1.8$. It can be found that paper co-authorship network has scale-free structure.

4 Conclusion and Future Works

In the paper, we first built the paper citation network from the bibliography lists of papers from the scientific papers database [6]. We then investigated the clustering coefficient, and the degree distribution by applying several network analysis techniques to the paper citation network. Consequently, we showed that the paper citation network has less grouped structure, and that the distribution of node in-degrees (i.e., the number of papers cited) follows a power-law. Moreover, in the same way, we built the paper co-authorship network based on the author information of papers. Consequently, we also show that the paper co-authorship network has grouped structure and scale-free structure.

As future work, developing a method for searching valuable nodes (papers or authors) using the same approach as [13] is of interest. Moreover, realization of data mining applications based on our findings on structure of the paper citation/co-authorship networks is also interesting.

Bibliography

[1] ALBERT, Reka, and Albert-Laszlo BARABASI, "Statistical mechanics of complex networks", *Reviews of Modern Physics* **74**, 47 (June 2002).

[2] BARABASI, A. L., H. JEONG, Z. NEDA, E. RAVASZ, A. SCHUBERT, and T. VIESEK, "Evolution of the social network of scientific collaborations", *PHYSICA A* **311** (Jan. 2002), 590–614.

[3] DAVIS, Gerald F., Mina YOO, and Wayne E. BAKER, "The small world of the american corporate elite, 1982-2001", *Strategic Organization* **1**, 3 (Aug. 2003), 301–326.

[4] GRANOVETTER, Mark, "The strength of weak ties: a network theory revisited", *Sociological Theory* **1** (Sept. 1983), 201–233.

[5] HOLGER, E., M. LUTZ-INGO, and B. STEFAN, "Scale-free topology of e-mail networks", *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)* **66**, 035103 (Sept. 2002), 1–4.

[6] IEEE COMMUNICATIONS SOCIETY, "Communications Engineering Technology A Comprehensive Collection of Papers 1952-2004", `http://www.comsoc.org/headlines/dvd.html`, 2005.

[7] JIN, M., Michelle GIRVAN, and M. E. J. NEWMAN, "The structure of growing social networks", *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)* **64**, 026118 (Oct. 2001), 381–399.

[8] LIBEN-NOWELL, David, and Jon KLEINVERG, "The link prediction problem for social networks", *Proceedings of the 12th ACM International Conference on Information and Knowledge Management (CIKM'03)*, (June 2003), 556–559.

[9] NEWMAN, M. E. J., "The structure of scientific collaboration networks", *Proc. Natl. Acad. Sci.* **98**, 2 (Jan. 2001), 404–409.

[10] NEWMAN, M. E. J., "The structure and function of complex networks", *SIAM Review* **45**, 2 (Mar. 2003), 167–256.

[11] PAGE, Lawrence, Sergey BRIN, Rajeev MOTWANI, and Terry WINOGRAD, "The pagerank citation ranking: Bringing order to the Web", *Tech. Rep. no.*, Stanford Digital Library Technologies Project, (1998).

[12] STOROGATZ, S. H., "Exploring complex networks", *Nature* **410**, 6825 (Mar. 2001), 268–276.

[13] SUGIYAMA, Kouhei, Osamu HONDA, Hiroyuki OHSAKI, and Makoto IMASE, "Application of network analysis techniques for Japanese corporate transaction network", *Proceedings of 6th Asia-Pacific Symposium on Information and Telecommunication Technologies (APSITT 2005)*, No. 9–10, (Nov. 2005), 387–392.

[14] WATTS, D. J., and S. H. STROGATZ, "Collective dynamics of 'small-world' networks", *Nature (London)* **393** (Oct. 1998), 440–442.