Comparative Study on Methods of Detecting Research Fronts Using Different Types of Citation

Naoki Shibata, Yuya Kajikawa, Yoshiyuki Takeda, and Katsumori Matsushima
Institute of Engineering Innovation, School of Engineering, University of Tokyo, 2-11-16 Yayoi, Bunkyo-ku, Tokyo 113-8656, Japan. E-mail: {shibata, kaji, takeda}@biz-model.t.u-tokyo.ac.jp; matsushima@iijmio-mail.jp

In this article, we performed a comparative study to investigate the performance of methods for detecting emerging research fronts. Three types of citation network, co-citation, bibliographic coupling, and direct citation, were tested in three research domains, gallium nitride (GaN), complex network (CNW), and carbon nanotube (CNT). Three types of citation network were constructed for each research domain, and the papers in those domains were divided into clusters to detect the research front. We evaluated the performance of each type of citation network in detecting a research front by using the following measures of papers in the cluster: visibility, measured by normalized cluster size, speed, measured by average publication year, and topological relevance, measured by density. Direct citation, which could detect large and young emerging clusters earlier, shows the best performance in detecting a research front, and co-citation shows the worst. Additionally, in direct citation networks, the clustering coefficient was the largest, which suggests that the content similarity of papers connected by direct citations is the greatest and that direct citation networks have the least risk of missing emerging research domains because core papers are included in the largest component.

Introduction

Historically, there have been a number of studies to detect emerging knowledge domains, also called research fronts, by analyzing the citation network of scientific publications. There are two main approaches. The first deals bibliometoric indicators such as Price's index (1970), which are time-based indicators, immediacy index (Garfield, 1972), and predictions of the future citation counts of individual papers. In recent studies, future times cited are predicted by current times cited (Adams, 2005), betweenness centrality, which is defined as the fraction of the shortest paths going through a given node (Chen, 2005; Leydesdorff, 2007a; Leydesdorff, 2007b), and a combination of these (Shibata, Kajikawa, & Matsushima, 2007). There are researches to extract factors to get Nobel Prize, such as time-interval between the discovery and its award, age of the discoverers and laureates in physics (Karazija & Momkauskaite, 2004), and nationality merit in physics, chemistry, and medicine (Braun, Szabadi-Peresztegi, & Kovacs-Nemeth, 2003). The other approach is to detect emerging clusters of densely connected papers. In his classic paper, de Solla Price (1965) originally introduced the concept of a research front, a research domain under development where papers cite each other densely. According to Price, there seems to be a tendency for scientists to cite the most recently published articles. A research front builds on recent work, and the network there becomes very tight. In a given field, a research front refers to the body of articles that scientists actively cite. Researchers have studied quantitative methods that can be used to identify and track a research front as it evolves over time. Small and Griffith (1974) represented currently activated scientific specialties as clusters of co-cited articles. Braam, Moed, & van Raan (1991) investigated the topics discussed in co-cited clusters by analyzing the frequency of indexing terms and classification codes occurring in these publications (Braam et al.). Besselaar and Leydesdorff (1996) showed mapping change and emergence in artificial intelligence domain.

Although there is general agreement to analyze citation patterns to detect emerging research domains, the type of "citation" differs among researches. There are three definitions of citation (Small, 1997): direct citation, co-citation (Small, 1973), and bibliographic coupling (Kessler, 1963). Co-citation is defined as the edge between two documents cited by the same paper(s). Bibliographic coupling is defined as the edge between two documents citing the same paper(s). If both paper A and B are cited by C, there is co-citation between A and B. And if both D and E cite F, there is bibliographic coupling between D and E. Small presented a method of tracking and predicting growth areas in the sciences by co-citation analysis that analyzed co-citation networks generated from the top 1% of highly cited papers (Small, 2006). Schiminovich (1971) classified academic

Received July 21, 2008; revised September 25, 2008; accepted October 11, 2008

 • Published online 18 December 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20994

publications automatically with recursive bibliographic coupling. Rousseau extracted significant substructures in citation networks with co-citation and bibliographic coupling (Fang & Rousseau, 2001; Egghe & Rousseau, 2002). Garfield (2004) generated historiographs mapping of knowledge domain of direct citation networks. Shibata Kajikawa, Takeda, and Matsushima (2008) proposed a method of detecting emerging domains with analysis of direct citation. But there is less research investigating the effect of citation type on performance in detecting research fronts. Recently, Klavans, and Boyack (2006) compared the performance of clustering in journal citation networks created by direct citation and co-citation. Their results suggested that a network of direct citation has higher content similarity. The purpose of this article is to study the characteristics of paper-paper citation networks created by different types of citation as well as their performance in the detection of research fronts.

This article studies the following three research domains. Gallium nitride (GaN) is widely recognized as a recent prominent innovation in the fields of applied physics and material science. Complex network (CNW) analysis is also recognized as pioneering a new research field. Nano-carbon (carbon nanotube [CNT]) is also widely recognized as a recent prominent innovation in the fields of applied physics and material science. We constructed the three types of citation network for each domain and divided the citation networks of each research domain into clusters to detect research fronts. We evaluated the performance of each method in detecting a research front by comparing the *visibility,* as measured by the normalized cluster size, *speed,* as measured by average publication year, and *topological relevance,* as measured by density, of the clusters. By considering the differences, we discuss which type of citation is most suitable for detecting emerging knowledge domains.

Overview of Research Domains

GaN, CNW, and CNT are typical examples of recent remarkable innovations having somewhat different characteristics. As explained later, research in GaN has incrementally developed in the field of applied physics. However, branching innovation (Shibata et al., 2008) occurs in CNW and CNT.

Within a very short period following the mid 1990s, researchers realized applications of GaN as blue and green light-emitting diodes (LEDs) and ultra violet (UV) and blue laser diodes (LDs). These products are now commercially available. Innovation in this research field motivates researchers to engage in and open huge new markets for manufacturers and customers. Some papers written by a researcher who worked in a Japanese firm has opened a new route to synthesizing high-quality GaN films having superior optical properties.

The second innovation is CNW, which was recently recognized as a new research field. Previously, CNWs have been researched in the following types of research: graph theory in mathematics, social network analysis in sociology, and applied physics. A prominent breakthrough occurred in the last domain, applied physics. Therefore, it can be expected that CNW research in applied physics forms a research front. An historical overview of these two domains, i.e., GaN and CNW, is in our previous paper (Shibata et al., 2007).

CNTs are useful in nanoscience and nanotechnology, due to superior electrical and mechanical properties. A CNT is a nano-sized carbon molecule having a morphology like a tube. Fullerenes are also a well-known nano-sized carbon material having a morphology like a ball. The existence of fullerenes was known earlier than that of nanotubes. But after the discovery of the carbon nanotube, the focus of researchers shifted fullerenes to nanotubes. Therefore, if we can detect research fronts that include papers where the discovery of the nanotube is mentioned, we might expect such shift of research focus earlier than competitors. In all of the above cases, earlier detection of research fronts is essential information for both researchers and research and development (R&D) managers to plan their research focus and strategy.

Research Methodology

The first step is to collect the data of each knowledge domain and to make citation networks. Citation networks were constructed by direct citation, co-citation, and bibliographic coupling. After constructing the networks, maximum connected components were extracted from each network. After extracting the maximum components, we divided the papers in the network into clusters. Finally, we evaluated the visibility, defined as normalized size, speed, defined as average publication year, and topological relevance, defined as density, of the clusters to which selected core papers belong. A list of core papers in each domain, which opened a new research frontier, is shown in Table 1.

Data Collection

We collected citation data from the Science Citation Index (SCI) and the Social Sciences Citation Index (SSCI) compiled by the Institute for Scientific Information (ISI), which maintains citation databases covering thousands of academic journals and offers bibliographic database services, because SCI and SSCI are two of the best sources for citation data. We used the Web of Science, which is a Web-based user interface

TABLE 1. Core papers that opened a new research frontier in three domains.

Research domain	Core papers
Gallium nitride	(a-1)NAKAMURA S, 1991, JPN J APPL PHYS PT 2, V30, P1705
	(a-2)NAKAMURA S, 1992, JPN J APPL PHYS PT 1, V31, P1258
	(a-3)NAKAMURA S, 1994, APPL PHYS LETT, V64, P1687
Complex networks	(b-1)Watts DJ, 1998, NATURE, V393, P440
	(b-2)Barabasi AL, 1999, SCIENCE, V286, P509
Carbon nanotube	(c-1)IIJIMA, S, 1991, NATURE, V354, P56

DOI: 10.1002/asi

of the ISI's citation databases. We searched the papers using the following terms as queries: "GaN OR gallium nitride" for the first domain, "social networks OR social network OR random networks OR random network OR small-world OR scale-free OR complex networks" for the second domain, and "carbon AND (nano* OR micro*)" for the third domain. In our method, queries were selected according to the following two steps: (a) the representative keyword, such as gallium nitride and social network, is selected and (b) if the definition of its domain is unclear, more keywords, such as random network, small-world, scale-free, and complex networks, were added. The second step is called query expansion (Kostoff, Eberhart, & Toothman, 1997). Our intention in using so many terms is to retain wide coverage of citation data in order to avoid omission of core papers.

The ISI's citation databases enable us to obtain both the attribute data of each paper such as the year published, title, author(s), abstract, and citation data. We create three types of citation network by regarding papers as nodes and three definitions of citations as edges, as shown in Figure 1. The network created in each year enables a time-series analysis of citation networks. When we create three types of citation network on year y, we use the data of papers published from 1970 to y, which are available on year y. In network analysis, only the data of the largest-graph component is used because this paper focuses on the relationship among papers, and we should therefore eliminate papers that have no link with any other papers. The number of papers contained in the largest component is 13,976 in GaN for 2004, 3,510 in CNW for 2004, and 23,459 in CNT for 2000.

Methods

With each citation network, we can calculate topological measures such as the number of nodes and edges. Topological clustering divides papers into clusters. Visibility (size normalized by the size of the largest component), speed (average publication year), and topological relevance (density) are determined after clustering for each cluster. We focus on topological clustering in order to discover tightly knit clusters with a high density of within-cluster edges, which enables the creation of a non-weighted graph consisting of a large number of nodes. Citation networks where each paper is connected by citation are divided into clusters. Although clustering has been difficult to achieve due to the difficulty in cluster analysis of non-weighted graphs with many nodes, there has in recent years been progress in methods of topological clustering (Newman, 2004).

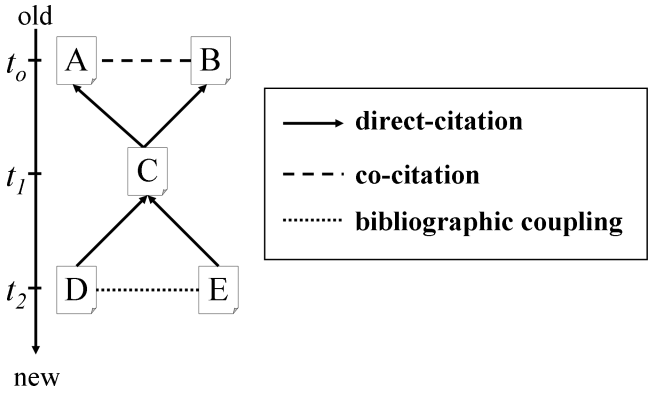


FIG. 1. Three types of citation.

Among many clustering methods and algorithms, in this article, we applied a method proposed by Newman (2004), which can deal with large networks in a relatively small calculation time in the order of $O((m+n)n)$, *or* $O(n^2)$ on a sparse network, with m edges and n nodes; therefore, this could be applied to large-scale networks Newman, (2004). The algorithm proposed is based on the idea of modularity. Modularity Q was defined as follows:

$$Q = \sum_s (e_{ss} - a_s^2) = Tr(e) - \|e\|^2 \qquad (1)$$

where e_{st} is the fraction of the edges in the network that connect nodes in cluster s to those in cluster t, and $a_s = \sum_t e_{st}$. The first part of the equation, $Tr(e)$, represents the sum of density of edges within each cluster. A high value of this parameter means that nodes are densely connected within each cluster. However, the maximum value of this ($Tr(e) = 1$) is given if whole nodes are regarded as one cluster. The second part of the equation, $\|e\|^2$, represents the sum of density of edges within each cluster when all edges are placed randomly. For instance, suppose that all nodes are regarded as one cluster. In this case, matrix e has only one row, $e = (1)$. Therefore, $Tr(e) = 1$, $\|e\|^2 = 0$, and $Q = 1$, which is the theoretical maximum value of Q. That is, Q is the fraction of edges that fall within communities, minus the expected value of the same quantity if the edges fall at random without regard for the community structure. Newman's method cuts off edges that connect clusters sparsely and extract clusters within which nodes are connected densely. A high value of Q represents good community division where only dense edges remain within clusters and sparse edges between clusters are cut off, and $Q = 0$ means that a particular division gives no more within-community edges than would be expected by random chance. Then, the algorithm to optimize Q over all possible divisions to find the best structure of clusters is as follows. Starting with a state in which each node is the only member of one of the n clusters, we repeatedly join clusters together in pairs, choosing at each step the join that results in the greatest increase in Q. The change in Q upon joining two clusters is given by

$$\Delta Q = e_{st} + e_{ts} - 2a_s a_t = 2(e_{st} - a_s a_t) \qquad (2)$$

In this article, we stop joining when ΔQ became negative because the purpose here is not to gain a whole dendrogram but extract more relevant structures in regard to citation networks.

Finally, we evaluate each citation method after clustering by measuring the visibility, speed, and topological relevance of the clusters to which these selected core papers belong.

Our basic concept is that the best type of citation is the one that can detect a larger and denser cluster at an earlier stage.

DOI: 10.1002/asi

When the normalized size of the cluster is larger, it means that the cluster is visible, and we can more easily distinguish the existence of emerging clusters from other clusters. When we have a young average publication year, it means that the cluster can be speedily detected, and we can earlier detect emerging clusters having core papers in them. If the cluster is denser, the papers in the cluster are topologically relevant, and we can check whether clustering is successful for dividing into clusters. Therefore, it seems reasonable to evaluate each linking method by these three measures.

Because the node size in the networks created by different types of citation differs, as shown later, the size is normalized to the relative size in order to compare the three types of citation, by $\frac{N(C)}{N}$, where N is the total number of entire nodes and $N(C)$ is the number of nodes in cluster C. The purpose of this normalization is to compare the relative size of clusters among three types of citations. The density is defined as $\frac{E_{\text{int}ra}(C)}{E(C)}$, where $E(C)$ is the number of edges from the nodes in cluster C, and $E_{\text{int}ra}(C)$ is the number of edges; both of the nodes are in cluster C. We can obtain different clustering results in each corpus according to the network constructed by the different three types of citation. We evaluate each method by comparing the normalized size and the average publication year of the cluster to which selected core papers in each corpus belong.

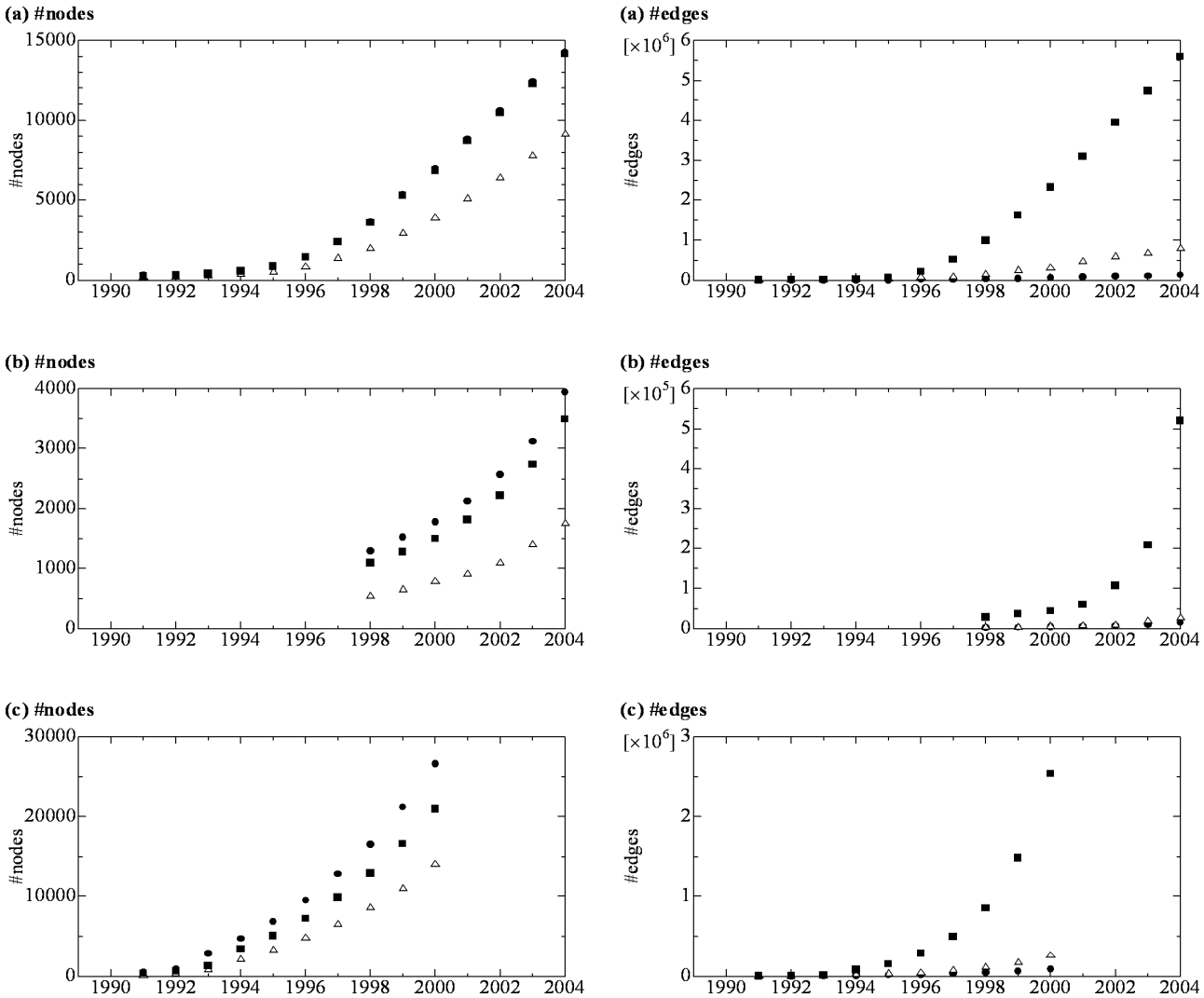
Results

First, basic characteristics of each network are shown in order to compare the differences among three types of citation networks. Secondary, results comparing performances in detecting emerging domains are shown.

Basic Characteristics of the Networks

As basic characteristics of networks, the number of nodes and edges and performance of each method are shown.

The number of nodes and edges in the networks. Figure 2 shows time series data on the number of nodes and edges of



Note. In each graph, the black circle shows direct citation, the white triangle shows co-citation, and the black square shows bibliographic coupling.

FIG. 2. The number of papers and edges of each domain: (a) gallium nitride, (b) complex networks, and (c) carbon nanotubes.

DOI: 10.1002/asi

each research domain. The number of nodes for direct citation was the largest and that for co-citation was the smallest in every type of citation and every year. The number of nodes differs among the three types of network because only the largest connected component is used. As in the co-citation (bibliographic coupling) network, some papers are isolated and not included in the largest component because these papers do not tend to be cited by (or cite) common papers.

The number of edges in bibliographic coupling was the largest and that in direct citation was the smallest. Normally, the number of the edges in co-citation and bibliographic coupling networks is larger than that in a direct citation network. Suppose that a paper that has k edges in direct citation is added to a network. With this addition, although the number of edges in direct citation increases by k, the number of edges in co-citation and bibliographic coupling increases by ${}_kC_2 \sim O(k^2)$. The result that the number of edges in bibliographic coupling was much larger than that in co-citation means that many papers tend to be cited by common papers rather than to cite common papers. Considering the above general trend in the number of nodes and edges, the densities of the citation networks defined by co-citation and bibliographic coupling are expected to be higher than those by direct citation.

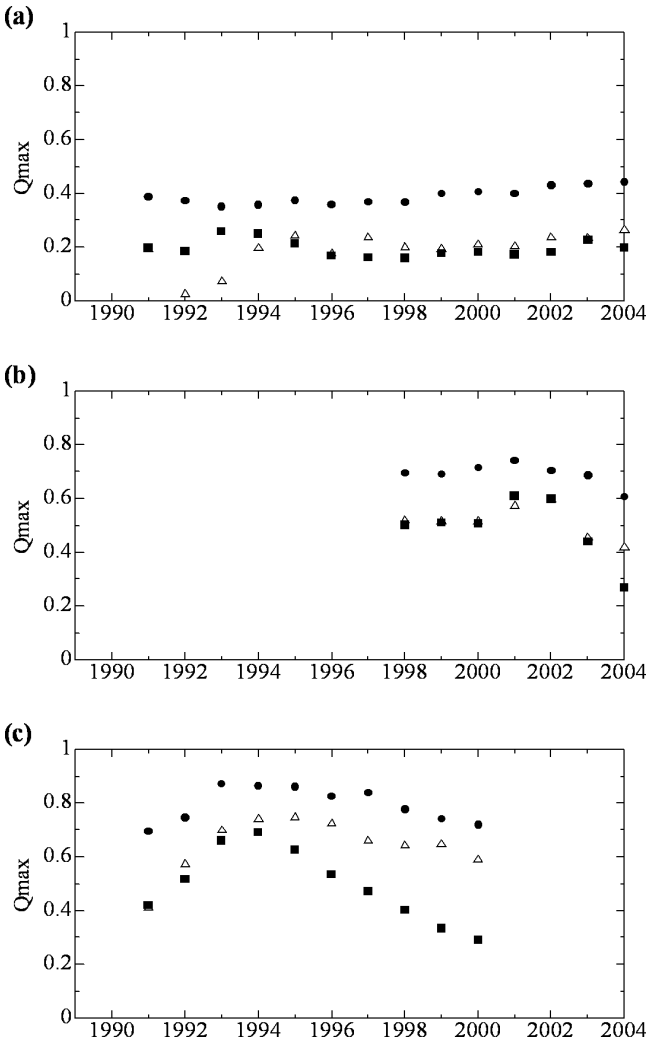
Modularity in the networks. Figure 3 shows the time series of *Qmax* of each research domain. In all years, *Qmax* in direct citation was the largest. These results are common regardless of the domain and mean that direct citation has a "locally dense and globally sparse" structure and can be divided into clusters better than the other two. In the bibliographic coupling network, the *Qmax* becomes smaller as the domain grows. This suggests that the bibliographic coupling network becomes random as the domain evolves, partly because it becomes denser not only locally but also globally and cannot be divided well. *Qmax* becomes higher when extracted clusters do not depend on each other, in other words, there are many intra-links but fewer inter-links. The low value of *Qmax* means that the network is close to a random network. Therefore, the above results indicate that networks defined by co-citation and bibliographic coupling are more random than those by direct citation.

Performance of Each Method in Detecting Emerging Domains

After clustering the networks, we evaluated the performance of the results in each citation network in detecting emerging research domains. The following domains, to which selected core papers in each domain belong, were tracked: visibility (as normalized size), speed (as average publication year), and topological relevance (as density). The normalized size, average publication year, and density of the clusters to which core papers belong are shown in Table 2.

Galium nitride (GaN). In the case of "NAKAMURA S, 1991, JPN J APPL PHYS PT 2, V30, P1705," the normalized size of clusters was the largest in bibliographic coupling and the smallest in direct citation, and the average publication year was the largest in direct citation and smallest in co-citation. Regarding "NAKAMURA S, 1992, JPN J APPL PHYS PT 1, V31, P1258," the normalized size of clusters was the largest in co-citation and the smallest in direct citation, and the average publication year was largest in direct citation and smallest in co-citation. In the case of "NAKAMURA S, 1994, APPL PHYS LETT, V64, P1687," the normalized size of clusters was largest in co-citation and smallest in direct citation and the average publication year was largest in bibliographic coupling and smallest in co-citation. Therefore, direct citation and bibliographic coupling seem to have superior performance in detecting research fronts including these core papers as early and effectively as possible because of their largest normalized size and younger age of clusters.

In the case of "NAKAMURA S, 1991, JPN J APPL PHYS PT 2, V30, P1705," the density in direct citation is largest and smallest in co-citation. Regarding "NAKAMURA S, 1992, JPN J APPL PHYS PT 1, V31, P1258," the density is largest in



Note. In each graph, the black circle shows direct citation, the white triangle shows co-citation, and the black square shows bibliographic coupling.

FIG. 3. *Qmax* value of each domain: (a) gallium nitride, (b) complex networks, and (c) carbon nanotubes.

DOI: 10.1002/asi

TABLE 2. Normalized size, average publication year, and density of the clusters to which core papers belong.

	Direct citation			Co-citation			Bibliographic coupling		
Year	Size	Average birth year	Density	Size	Average birth year	Density	Size	Average birth year	Density
(a-1)									
1991	0.18	1989.00	73%	–	–	–	0.45	1985.14	60%
1992	0.22	1990.18	69%	–	–	–	0.50	1984.77	68%
1993	0.27	1991.36	68%	–	–	–	0.44	1991.15	69%
1994	0.28	1992.18	70%	–	–	–	0.43	1992.85	72%
1995	0.36	1993.72	75%	–	–	–	0.11	1992.81	36%
1996	0.32	1994.97	68%	–	–	–	0.50	1989.35	63%
1997	0.30	1995.99	67%	–	–	–	0.47	1991.82	60%
1998	0.32	1992.64	70%	–	–	–	0.33	1996.71	57%
1999	0.35	1994.45	75%	–	–	–	0.68	1996.30	70%
2000	0.33	1995.39	74%	–	–	–	0.67	1997.16	72%
2001	0.33	1996.03	75%	–	–	–	0.54	1999.13	64%
2002	0.28	1999.72	68%	–	–	–	0.55	1999.77	66%
2003	0.34	1997.42	77%	–	–	–	0.33	1999.19	69%
2004	0.33	2000.69	75%	–	–	–	0.07	1998.91	40%
(a-2)									
1992	0.22	1990.18	69%	0.70	1981.61	72%	0.50	1984.77	68%
1993	0.27	1991.36	68%	0.42	1989.91	40%	0.44	1991.15	69%
1994	0.28	1992.18	70%	0.56	1991.66	67%	0.43	1992.85	72%
1995	0.36	1993.72	75%	0.65	1993.02	75%	0.11	1992.81	36%
1996	0.32	1994.97	68%	0.43	1987.63	71%	0.50	1989.35	63%
1997	0.30	1995.99	67%	0.60	1994.91	64%	0.47	1991.82	60%
1998	0.12	1996.86	52%	0.44	1995.04	65%	0.33	1996.71	57%
1999	0.35	1997.86	72%	0.28	1996.25	60%	0.68	1996.30	70%
2000	0.32	1998.28	71%	0.47	1997.94	63%	0.67	1997.16	72%
2001	0.33	1996.03	75%	0.35	1994.62	70%	0.54	1999.13	64%
2002	0.31	1999.63	77%	0.18	1998.58	50%	0.55	1999.77	66%
2003	0.20	2000.12	74%	0.51	1999.41	64%	0.33	1999.19	69%
2004	0.31	1998.61	76%	0.56	2000.21	68%	0.07	1998.91	40%
(a-3)									
1994	0.28	1992.18	70%	0.56	1991.66	67%	0.43	1992.85	72%
1995	0.36	1993.72	75%	0.65	1993.02	75%	0.48	1994.22	75%
1996	0.32	1994.97	68%	0.43	1987.63	71%	0.49	1995.42	77%
1997	0.30	1995.99	67%	0.60	1994.91	64%	0.47	1991.82	60%
1998	0.22	1996.82	65%	0.24	1996.59	32%	0.33	1996.71	57%
1999	0.21	1997.39	68%	0.48	1995.79	64%	0.16	1996.76	61%
2000	0.23	1998.33	66%	0.47	1997.94	63%	0.21	1997.44	66%
2001	0.33	1996.03	75%	0.35	1994.62	70%	0.38	1996.64	71%
2002	0.28	1999.72	68%	0.45	1997.09	64%	0.36	1997.34	73%
2003	0.09	2000.50	55%	0.42	1997.26	66%	0.33	1999.19	69%
2004	0.31	1998.61	76%	0.56	2000.21	68%	0.34	1998.21	70%
(b-1)									
1998	–	–	–	–	–	–	–	–	–
1999	0.01	1998.72	98%	0.02	1998.50	95%	–	–	–
2000	0.04	1999.56	99%	0.05	1998.68	99%	–	–	–
2001	0.09	2000.21	99%	0.10	1999.09	97%	–	–	–
2002	0.17	2001.15	99%	0.30	1993.83	81%	–	–	–
2003	0.22	2002.00	99%	0.22	2000.67	95%	–	–	–
2004	0.30	2002.88	98%	0.31	2001.71	95%	–	–	–
(b-2)									
1999	0.01	1998.72	98%	–	–	–	0.02	1998.80	97%
2000	0.04	1999.56	99%	0.05	1998.68	99%	0.06	1999.18	99%
2001	0.09	2000.21	99%	0.10	1999.09	97%	0.12	1999.85	99%
2002	0.17	2001.15	99%	0.30	1993.83	81%	0.21	2001.24	98%
2003	0.22	2002.00	99%	0.22	2000.67	95%	0.28	2002.05	98%
2004	0.30	2002.88	98%	0.31	2001.71	95%	0.08	2002.21	31%

(*Continued*)

DOI: 10.1002/asi

TABLE 2. (*Continued*)

	Direct citation			Co-citation			Bibliographic coupling		
Year	Size	Average birth year	Density	Size	Average birth year	Density	Size	Average birth year	Density
(c-1)									
1991	0.10	1967.97	99%	–	–	–	0.26	1987.50	97%
1992	0.08	1991.60	98%	0.27	1984.36	95%	0.10	1991.91	95%
1993	0.10	1992.56	97%	0.15	1991.49	96%	0.12	1992.64	97%
1994	0.09	1993.19	98%	0.13	1988.63	91%	0.11	1993.32	99%
1995	0.10	1993.83	99%	0.13	1992.01	98%	0.11	1994.07	97%
1996	0.10	1994.54	98%	0.10	1990.51	84%	0.13	1994.68	98%
1997	0.11	1995.25	98%	0.13	1993.95	96%	0.14	1995.34	99%
1998	0.11	1996.05	98%	0.32	1992.88	87%	0.14	1996.23	98%
1999	0.14	1996.92	98%	0.24	1995.49	98%	0.01	1994.43	55%
2000	0.21	1996.98	97%	0.26	1996.38	98%	0.13	1997.49	90%

co-citation and smallest in bibliographic coupling. In the case of "NAKAMURA S, 1994, APPL PHYS LETT, V64, P1687," the density is largest in bibliographic coupling and smallest in co-citation. Looking at the results of density and *Qmax* in the above, from the viewpoint of clustering performance, direct citation could extract densely connected clusters better than bibliographic coupling.

Complex networks (CNWs). In the case of "Watts DJ, 1998, NATURE, V393, P440," the normalized size of clusters in co-citation was slightly larger than in direct citation, and the average publication year in direct citation was slightly larger than in co-citation. This paper was not involved in the largest component of bibliographic coupling and could not be detected. Regarding "Barabasi AL, 1999, SCIENCE, V286, P509," the normalized size of clusters in bibliographic coupling was slightly larger than in direct citation, and the average publication year in direct citation was slightly larger than in bibliographic coupling. This paper was not involved in the largest component of co-citation and could not be detected in 1999. Therefore, direct citation seems to have superior performance in detecting research fronts including these core papers as early and effectively as possible because of its largest normalized size and younger age of clusters.

In the case of "Watts DJ, 1998, NATURE, V393, P440," the density in direct citation is slightly larger than in co-citation. Regarding "Barabasi AL, 1999, SCIENCE, V286, P509," the density in direct citation is slightly larger than in bibliographic coupling. Looking at the results of density and *Qmax* in the above, from the viewpoint of clustering performance, direct citation could extract densely connected clusters better than the other two.

Carbon nanotubes (CNTs). In the case of "IIJIMA, S, 1991, NATURE, V354, P56," the normalized size of clusters in bibliographic coupling was larger than in direct citation, and the average publication year in bibliographic coupling was larger than in direct citation. This paper was not involved in the largest component of co-citation and could not be detected in 1991. Therefore, bibliographic coupling seems to have superior performance in detecting research fronts including these core papers as early and effectively as possible because of its largest normalized size and younger age of clusters.

In the case of "IIJIMA, S, 1991, NATURE, V354, P56," the density in direct citation is larger than in bibliographic coupling. Looking at the results of density and *Qmax* in the above, from the viewpoint of clustering performance, direct citation could extract densely connected clusters better than the other two.

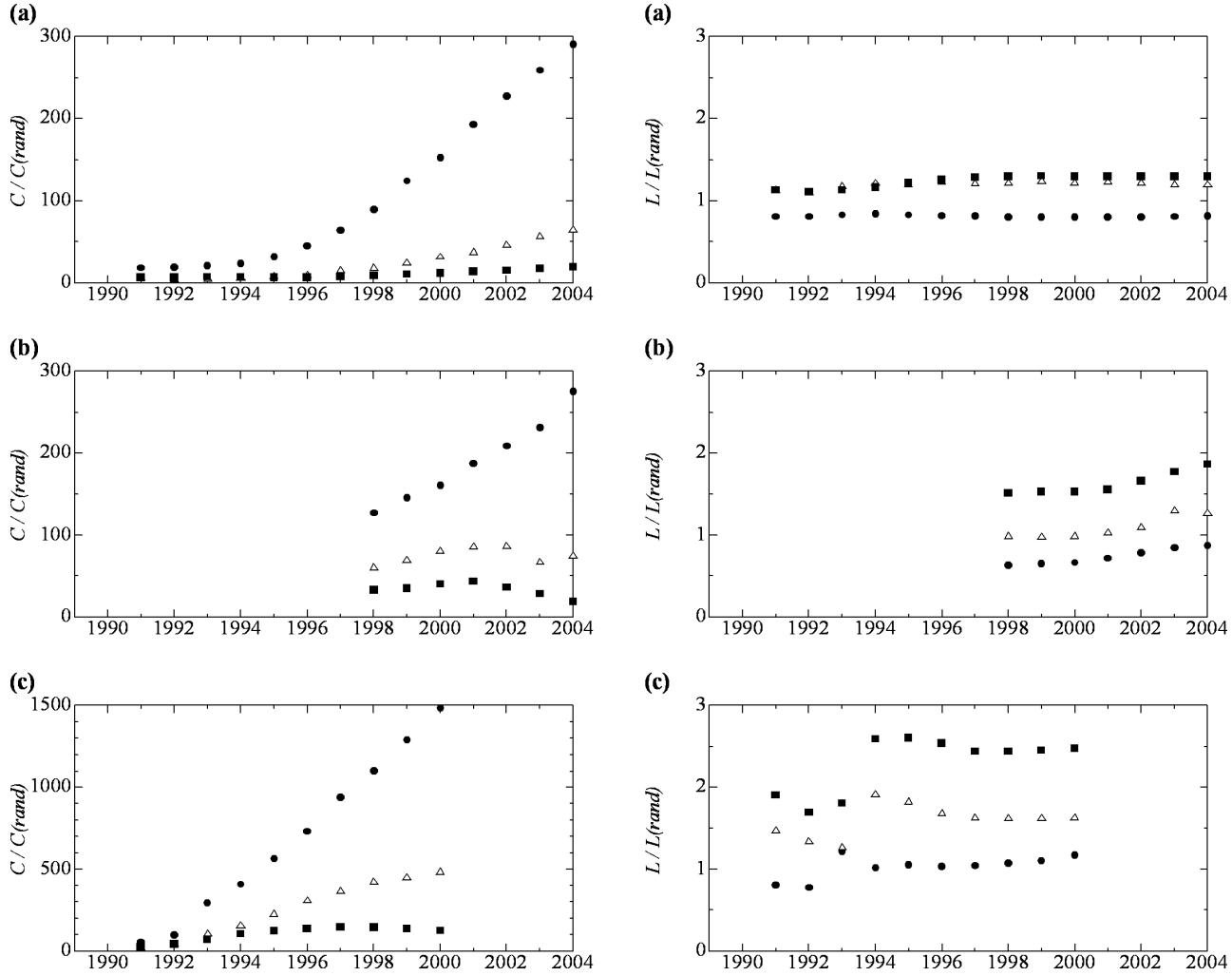
Discussions

A comparison of the results is shown in Table 3. Co-citation was the worst because of the existence of a time lag in co-citation as pointed out by Hopcroft, Khan,

TABLE 3. Brief result of comparison of three types of citation.

Research domain	Core paper	Visibility (normalized size)	Speed (average birth year)	Topological relevance (density)
Gallium nitride	(a-1) NAKAMURA S, 1991, JPN J APPL PHYS PT 2, V30, P1705	biblio > co > direct	direct > biblio > co	direct > biblio > co
	(a-2) NAKAMURA S, 1992, JPN J APPL PHYS PT 1, V31, P1258	biblio > direct > co	direct > co > biblio	direct > biblio > co
	(a-3) NAKAMURA S, 1994, APPL PHYS LETT, V64, P1687	co > biblio > direct	biblio > co > direct	biblio > direct > co
Complex networks	(b-1) Watts DJ, 1998, NATURE, V393, P440	direct ≒ co	direct ≒ co	direct > co
	(b-2) Barabasi AL, 1999, SCIENCE, V286, P509	direct ≒ biblio	direct ≒ biblio	direct > biblio
Carbon nanotube	(c-1) IIJIMA, S, 1991, NATURE, V354, P56	biblio > direct	biblio > direct	direct > biblio

DOI: 10.1002/asi



Note. In each graph, the black circle shows direct citation, the white triangle shows co-citation, and the black square shows bibliographic coupling.

FIG. 4. Normalized characteristic path length and clustering coefficient of each network: (a) gallium nitride, (b) complex networks, and (c) carbon nanotubes.

Kulis, & Selman (2004). In co-citation, two papers are linked if they are both cited by another paper published later. A time lag is needed to build up a citation record for co-citation.

There is no lag in bibliographic coupling either. Direct citation and bibliographic coupling analysis are more sensitive to recent citations than co-citation analysis. In this point, bibliographic coupling could be expected to be best because it could potentially detect more edges earlier than the other two methods. However, our results did not support this hypothesis.

As an alternative potential explanation, we evaluate the cohesiveness of citations in each cluster by small-world properties. How "small world" a network is based upon characteristic path length L and clustering coefficient C (Watts & Strogatz, 1998). The characteristic path length L is defined as the number of edges in the shortest path between two nodes, averaged over all pairs of nodes. The clustering coefficient C is defined as follows. Suppose that node n has k neighbors; then, at most, ${}_kC_2 = \frac{k(k-1)}{2}$ edges can exist between them (this occurs when every neighbor of n is connected to every other neighbor of n). Let $C(n)$ denote the fraction of these allowable edges that actually exist. Define C as the average of $C(n)$ over all n. Figure 4 shows the normalized values of clustering coefficient C and characteristic path length L of each network. These values are normalized by *C(rand)* and *L(rand)*, which could be calculated as $C(rand) \approx \frac{\langle k \rangle}{N}$ and $L(rand) \approx \frac{\ln(N)}{\ln(\langle k \rangle)}$, where N is the number of nodes and $<k>$ is the number of average edges per node. In all these networks, because C is much larger than *C(rand)* and L is close to *L(rand)*, the networks have the "small-world" characteristics. In a network with large C, node X's neighbor's neighbor tends to be a neighbor of node X. Comparing the three types of citation, *C/C(rand)* was the largest in direct citation and the smallest in bibliographic coupling. *L/L(rand)* was the largest in bibliographic coupling and the smallest in direct citation. The result that the *C/C(rand)* value in direct citation was the largest means that the edges in direct citation are more cohesive than in the other two, matching the above results that density in emerging clusters tends to be high in direct citation networks. On the other hand, the numbers of edges in

DOI: 10.1002/asi

co-citation and bibliographic coupling are much greater than in direct citation networks and, regarding the small value of *C/C(rand),* co-citation and bibliographic coupling could be regarded as "random networks." The results of *Qmax* shown in Figure 3 also accord with these results. In summary, direct citation networks tend to be the most cohesive and seem to connect content-similar papers. With this feature, direct citation should perform the best in detecting research fronts including core papers as early and effectively as possible.

One more result supports the hypothesis that direct citation is better than bibliographic coupling. Bibliographic coupling and co-citation sometimes failed to include core papers in the largest component, whereas direct citation did not. For instance, with bibliographic coupling, the core paper "Watts DJ, 1998, NATURE, V393, P440" was not in the largest component as shown in Table 2. The results show that there is a risk of missing emerging research domains with bibliographic coupling and co-citation because core papers may not be included in the largest component.

It is possible that bibliographic coupling was slightly worse than direct citation because of the process of creating networks. In our method, only the citations among papers that were collected by the queries were used to create each citation network. We performed an additional examination to create one more network, named bibliographic coupling(ex), using the data of "one path" from the collected papers. In the process of creating bibliographic coupling(ex) network, papers cited by collected papers are considered but not included as nodes. In the above case, we can make a link between A and B even if A and B are directly linked and X is not included in the network as a node. It is possible that bibliographic coupling(ex) networks show superior performance by considering this additional link. The results indicated that the bibliographic coupling(ex) network could not extract emerging clusters well. The bibliographic coupling(ex) network was so dense that this network was close to a random network with small *Qmax* and low density of emerging clusters.

Conclusion

This article reports a comparative study to investigate the performance of methods for detecting emerging research fronts among co-citation, bibliographic coupling, and direct citation networks. A case study in three research domains, gallium nitride, complex networks, and carbon nanotubes, was performed. After three types of citation networks were constructed, papers in each research domain were divided into clusters using a topological clustering. We evaluated the visibility, defined as normalized size, speed, defined as average publication year, and topological relevance, defined as density, of the clusters to which selected core papers belong.

The best data is direct citation, which could detect large emerging clusters earlier, and co-citation is the worst. The reason that co-citation is the worst is that a time lag is inescapably needed in order for papers to build up a co-citation record. Comparing direct citation and bibliographic coupling, the clustering coefficient in direct citation is the largest, which suggests that the content similarity of papers connected by direct citations is the greatest, and the direct citation network has the least risk of missing emerging research domains because core papers are always included in the largest component.

One of the potential weaknesses of citation analysis to detect emerging research front is a time lag to cite (or be cited). Although, in this article, we analyzed only topological data, semantic similarity analysis based on textual data may have the potential to detect emerging research fronts earlier and more precisely. Further work is necessary to compare the performance of a link-based approach, text-based approach, and hybrid approach to detecting emerging research fronts.

References

Adams, J. (2005). Early citation counts correlate with accumulated impact. Scientometrics, 63(3), 567–581.

Besselaar, van den P., & Leydesdorff, L. (1996). Mapping change in scientific specialties: A scientometric reconstruction of the development of artificial intelligence. Journal of the American Society for Information Science, 47, 415–436.

Braam, R.R., Moed, H.F., & van Raan, A.F.J. (1991). Mapping of science by combined co-citation and word analysis. i. structural aspects. Journal of the American Society for Information Science, 42, 233–251.

Braun, T., Szabadi-Peresztegi, Z., & Kovacs-Nemeth, E. (2003). No-bells for ambiguous lists of ranked Nobelists as science indicators of national merit in physics, chemistry and medicine, 1901–2001. Scientometrics, 56(1), 3–42.

Chen, C. (2005). Measuring the movement of a research paradigm. Proceedings of SPIE-IS&T: Visualization and Data Analysis 2005 (VDA2005) (pp. 63–76). San Jose: SPIE and IS&T.

de Solla Price, D.J. (1965). Networks of scientific papers. Science, 149, 510–515.

Egghe, L., & Rousseau, R. (2002). Co-citation, bibliographic coupling and a characterization of lattice citation networks. Scientometrics, 55(3), 349–361.

Fang, Y., & Rousseau, R. (2001). Lattices in citation networks: An investigation into the structure of citation graphs. Scientometrics, 50(2), 273–287.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. Science, 178(4060), 471–479.

Garfield, E. (2004). Historiographic mapping of knowledge domains literature. Journal of Information. Science, 30(2), 119–145.

Hopcroft, J., Khan, O., Kulis, B., & Selman, B. (2004). Tracking evolving communities in large linked networks. Proceedings of the National Academy of Sciences, 101, 5249–5253.

Karazija, R., & Momkauskaite, A. (2004). The Nobel prize in physics—Regularities and tendencies. Scientometrics, 61(2), 191–205.

Kessler, M.M. (1963). Bibliographic coupling between scientific papers. American Documentation, 14, 10–25.

Klavans, R., & Boyack, K. W. (2006). Identifying a better measure of relatedness for mapping science. Journal of the American Society for Information Science and Technology, 57, 251–263.

Kostoff, R.N., Eberhart, H.J., & Toothman, D.R. (1997). Database tomography for information retrieval. Journal of Information Science, 23, 301–311.

Leydesdorff, L. (2007a). "Betweenness centrality" as an indicator of the "interdisciplinarity" of scientific journals. Journal of the American Society for Information Science and Technology, 58(9), 1303–1309.

Leydesdorff, L. (2007b). Mapping interdisciplinarity at the interfaces between the Science Citation Index and the Social Science Citation Index. Scientometrics, 71(3), 391–405.

DOI: 10.1002/asi

Newman, M.E.J. (2004). Fast algorithm for detecting community structure in networks. Physical Review E, 69, 066133.

Schiminovich, S. (1971). Automatic classification and retrieval of documents by means of a bibliographic pattern discovery algorithm. Information Storage and Retrieval, 6, 417–435.

Shibata, N., Kajikawa, Y., & Matsushima, K. (2007). Topological analysis of citation networks to discover the future core articles. Journal of the American Society for Information Science and Technology, 58(6), 872–882.

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. Technovation, 28(11), 758–775.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science, 24, 265–269.

Small, H. (1997). Update on science mapping: Creating large document spaces. Scientometrics, 38, 275–293.

Small, H. (2006). Tracking and predicting growth areas in science. Scientometrics, 68(3), 595–610.

Small, H.G., & Griffith, B.C. (1974). The structure of scientific literatures: I. identifying and graphing specialties. Science Studies, 4, 17–40.

Watts, D.J., & Strogatz, S.H. (1998). Collective dynamics of "small-world" networks. Nature, 393, 440–442.

DOI: 10.1002/asi