This is the preliminary version of the accepted JASIST paper

Scholarly network similarities: How bibliographic coupling networks, citation networks, co-citation networks, topical networks, coauthorship networks, and co-word networks relate to each other

Erjia Yan[1], **Ying Ding**

School of Library and Information Science, Indiana University, Bloomington, USA

Abstract

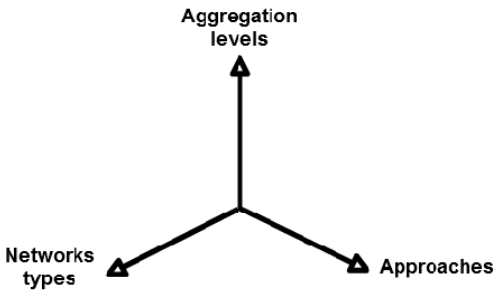
This study is motivated to explore the similarity among six types of scholarly networks aggregated at the institution level, including bibliographic coupling networks, citation networks, co-citation networks, topical networks, coauthorship networks, and co-word networks. Cosine distance is chosen to measure the similarities among the six networks. We find that topical networks and coauthorship networks have the lowest similarity; co-citation networks and citation networks have high similarity; bibliographic coupling networks and co-citation networks have high similarity; and co-word networks and topical networks have high similarity. In addition, through multidimensional scaling, two dimensions can be identified among the six networks: Dimension 1 can be interpreted as "citation-based vs. non-citation-based", and Dimension 2 can be interpreted as "social vs. cognitive". We recommend the use of hybrid or heterogeneous networks to study research interaction and scholarly communications.

Introduction

In recent years, we have witnessed a growing trend in the study of various types of scholarly networks, wherein a node usually denotes an academic entity, such as a paper, a journal, or an author, and a link usually denotes relationships such as citation, coauthorship, co-citation, bibliographic coupling, or co-word. Through scholarly network analysis, scientists and policy makers have gained unprecedented insights into the interaction of these research aggregates. Network-based bibliometric study in general can

[1] *Correspondence to*: Erjia Yan, School of Library and Information Science, Indiana University, 1320 E. 10th St., LI011, Bloomington, Indiana, 47405, USA. Email: eyan@indiana.edu

be presented in a three-dimensional framework that includes approaches, networks types, and aggregation levels (Figure 1).



Dimensions	
Aggregation levels	paper, author, journal, institution, field, etc.
Approaches	micro-level (centrality measures, PageRank and variants); meso-level (various clustering techniques); macro-level (degree distribution, density, etc.)
Networks	citation, co-citation, bibliographic coupling, coauthorship, co-word, etc

Figure 1. A 3-D presentation of network-based bibliometric studies

Approaches. The popularity of network studies, influenced by social studies of human interactions, was accelerated by the discovery of small-world and scale-free properties and enriched by various macro-level statistics, meso-level clustering techniques, and micro-level indicators. Approaches that scholars have used to examine scholarly networks can fit into three categories: macro-level statistics, meso-level techniques, and micro-level indicators. Macro-level statistics are useful in identifying the global structural features of networks. Meso-level approaches focus on the behavior of a group of actors, where various clustering techniques can be classified into this category. Micro-level indicators are useful in understanding individual node's power, stratification, ranking, and inequality in social structures (Wasserman & Faust, 1994).

Network types. In addition to the different approaches, the interaction of research aggregates can be explored from different types of scholarly networks. Each type of scholarly network has its own use and can bring a range of perspectives to the study of research interactions and scholarly communications. For example, social networks such as coauthorship networks focus on finding patterns of contacts or interactions between social actors. Similarity-based networks such as co-citation networks, bibliographic coupling networks, and co-word networks focus on identifying research topics or disciplines. In citation networks, each node is a piece of knowledge and a link denotes the knowledge flow.

Aggregation levels. In these network types mentioned above, an article is usually a single research unit that can be aggregated into several higher levels, for instance, the author unit, the journal unit, the institution unit, and the field unit. Through studies of different research aggregates, multiple focus lenses have been provided that allow us to zoom in and gain a concrete and detailed perspective on research interaction, while zooming out allows us to obtain a holistic and integrated view of the interacting institutions and disciplines. Previous efforts on scholarly network analyses have mainly emphasized lower research aggregates such as papers, authors, and journals. The findings

from higher-level analyses, however, can provide richer contexts to study scholarly communications. For example, institutional scholarly networks analysis provides an opportunity to combine mappings from social, geographical, and cognitive perspectives.

Previous network-based bibliometric studies usually chose one type of network at one aggregation level (e.g., author co-citation network) and used one approach (e.g., micro-level) to address certain research questions. The choice of network type can sometimes be inconsistent, and thus problematic. Boyack and Klavans (2010) pointed out that "[C]o-citation analysis was adopted as the de facto standard in the 1970s, and has enjoyed that position of preference ever since. There has been a recent resurgence in the use of bibliographic coupling that is challenging the historical preference for co-citation analysis" (p. 2390). This study helps provide a solution by identifying the similarity among six types of scholarly networks aggregated at the institution level, including bibliographic coupling networks, citation networks, co-citation networks, topical networks, coauthorship networks, and co-word networks. The results of this study can provide a better understanding of scholarly networks and contribute to social and cognitive studies of institution interactions.

Literature review

Scholarly networks: approaches, network types, and aggregation levels

In bibliometrics, scholarly networks have been largely explored at meso- and micro-levels. Bibliometricians and scientometricians have been dedicated to providing more accurate clustering and ranking approaches. Scholars working on meso-level scholarly network analysis have applied various clustering techniques to identify topics or map the backbone of science. Those methods include multidimensional scaling (e.g., White & McCain, 1998), k-means (e.g., Yan, Ding, & Jacob, 2012), and modularity-based clustering techniques (e.g., Van Eck, Waltman, Dekker, & Van den Berg, 2010). At the micro-level, scholars are seeking for more fine-grained bibliometric indicators to evaluate research progress. Simple citation counting has served as a formal instrument for quantitative scientific evaluation for several decades. Although it is easy to comprehend and implement, this tool does not take into account the linking structure of citing journals, citing authors, or citing articles. Noticing such problems, scholars have proposed various network-based bibliometric indicators that are able to consider the source of citation endorsement, such as Y-factor (Bollen, Rodriguez, & Van de Sompel, 2006), CiteRank (Walker, Xie, Yan, & Maslov, 2007), Eigenfactor (West, Bergstrom, & Bergstrom, 2010), and P-Rank (Yan, Ding, & Sugimoto, 2011).

Along with real connections (e.g., citation and coauthorship connections), some scholarly networks are constructed based on similarity connections. Compared to real connections, similarity-based connections are artificially made, such as the number of times two

authors were co-cited, or the number of times two words co-occurred. Co-occurrence networks are generally used to identify the research fields and study interdisciplinarities. Examples of co-occurrence networks are author co-citations networks (White & McCain, 1998), paper co-citation networks (Small, 1973), journal co-citation networks (Ding, Chowdhury, & Foo, 2000), and co-word relations (Milojević, Sugimoto, Yan, & Ding, 2011). As studies on these scholarly networks have habitually used only one type of network, their findings are discrete and cannot be used to answer a broader spectrum of questions regarding scholarly interactions.

In regards to aggregation levels, papers are the basic research unit, which can be aggregated into several higher research units, such as author unit, journal unit, institution unit, or country unit. At the paper level, scholarly networks usually utilize PageRank or its variants to differentiate the weight of citations according to the provenance of citation endorsements (e.g., Chen, Xie, Maslov, & Render, 2007; Ma, Guan, & Zhao, 2008; Waltman, Yan, & Van Eck, 2011). At the author level, Radicchi, Fortunato, Markines, and Vespignani (2009) constructed an author citation network to differentiate the scientific credits of authors based on the status of their citing authors. Attempts have also been made to differentiate popular and prestigious authors in coauthorship networks (Ding & Cronin, 2011; Yan & Ding, 2011). Journal citation networks are often employed to capture knowledge diffusion among research domains. The underlying assumption of indicators used in journal citation networks (such as Eigenfactor, Y-Factor, and SCImago Journal Rank Indicator) is that a journal is said to be prestigious if it is cited by other prestigious journals. Higher-level research aggregates have also been explored in recent years. For example, Yan and Sugimoto (2011) formulated a linear regression model to study the factors associated with institutional citation behaviors.

Hybrid and heterogeneous networks

There is a current trend in bibliometrics to use hybrid approaches in identifying research topics. Liu, Yu, Janssens, Glanzel, Moreau, and De Moor (2010) presented a framework of hybrid clustering to combine lexical and citation data for journal analysis. Zitt, Lelu, and Bassecoulard (2011) examined the convergence of two thematic mapping approaches, citation-based and word-based. They found these two approaches yield quite different outcomes and cannot be substituted with each other. Boyack and Klavans (2010) examined several types of scholarly networks, including a co-citation network, a bibliographic coupling network, and a citation network, in the interest of selecting the network that can best represent the research front in biomedicine. They used within-cluster textual coherence and grant-to-article linkage indexed by MEDLINE as accuracy measurements, and found the bibliographic coupling-based citation-text hybrid approach, which couples both references and words from title/abstract, outperforms other approaches. Janssens, Glänzel, and De Moor (2008) proposed a novel hybrid approach that integrates two types of information, citation (in the form of a term-by-document

matrix) and text (in the form of a cited_references-by-document matrix), through the use of Fisher's inverse chi-square method. This method can effectively combine matrices with different distributional characteristics. They found that the hybrid approach outperforms the text-only approaches by successfully assigning papers into correct clusters.

Hybrid approaches usually couple different types of networks in an intuitive way, without consideration of edge semantics. In addition to the efforts made on hybrid approaches, scholars have constructed heterogeneous scholarly networks that can incorporate different academic entities while keeping edge semantics. The study of heterogeneous networks has evolved from bi-typed networks (e.g., Zhou, Orshanskiy, Zha, & Giles, 2007; Sun, Yu, & Han, 2009; Sayyadi & Getoor, 2009; Yan, Ding, & Sugimoto, 2011) to star-typed heterogeneous networks (e.g., Sun, Barber, Gupta, Aggarwal, & Han, 2011). The co-ranking model (Zhou et al., 2007) coupled two networks, a coauthorship network and a paper citation network. FutureRank (Sayyadi & Getoor, 2009) used coauthorship and citation networks to predict future citations. P-Rank (Yan, Ding, & Sugimoto, 2011) differentiated the weight of each citation based on its citing papers, citing journal, and citing authors through a citation network and two authorship networks. Sun et al. (2011) defined a schema for bibliographic networks that contains four academic entities (paper, author, topic, and venues) and four relationships (citation, collaboration, publication, and mentioning). This schema can be used to predict coauthor relationships and to rank academic entities.

While the hybrid approach furnishes a sound starting point for ongoing studies on scholarly networks, simply assembling different networks may cause unexpected problems, as we are unaware of how different scholarly networks relate to each other. This study is therefore motivated to examine the similarity of six different scholarly networks and aims to advance the scholarship of scholarly network analyses.

Data

The dataset used in this analysis was drawn from all documents in the 59 journals indexed in the 2008 version of the Journal Citation Reports (JCR) under the Information Science & Library Science category[2]. All document types published within these journals from January 1965 to February 2010 were downloaded for analysis[3].

Data were processed in several steps. The first step was to filter the dataset in order to create a local citation network between institutions. The second step involved identifying

[2] There are 61 journals categorized as Information Science & Library Science in 2008; two journals written in foreign languages were excluded, PROF INFORM and Z BIBL BIBL, making the total number of journals in the data set 59.

[3] See data and visualizations at http://info.slis.indiana.edu/~eyan/papers/citation/

unique institution names from the affiliation data (see Yan and Sugimoto, 2011, for a detailed description of data processing).

The dataset was then divided into four subsets based on the citing papers' year of publication. The latest three time periods are selected. Time span is longer for the first period as the first years provided insufficient data to form comparable networks. Table 1 shows the size of institution citation networks and paper citation networks.

Institution collaboration (coauthor) networks, bibliographic coupling networks (BGcoupling), co-citation networks, co-word networks, and topical networks were constructed using the same vertex list.

Table 1. Size of institution citation networks

Time	Size of institution citation networks	Size of paper citation networks
1991-2000	2,906*2,906	9,750*9,750
2001-2005	3,010*3,010	9,280*9,280
2006-2010	3,783*3,783	10,998*10,998

The construction of citation and coauthorship networks

The dataset consists of documents with at least one author affiliation that had been cited by another document (containing author affiliations) within the dataset. Citation counts between documents were calculated, using the concept of "internal citation". That is, the number of times an article has been cited by other articles in the network (but not in the whole Web of Science database). An operationalized procedure is illustrated in Figure 2.

First, two basic matrices are constructed. One is the institution authorship network (W): W_{ij}=1 if institution j wrote paper i, and 0 otherwise. The other is the paper citation network (C): C_{ij}=1 if paper j is cited by paper i, and 0 otherwise. Based on the matrix manipulation rule, institution citation network can then be obtained by $W'*C*W$, and institution coauthorship network can be obtained by $W'*W$.

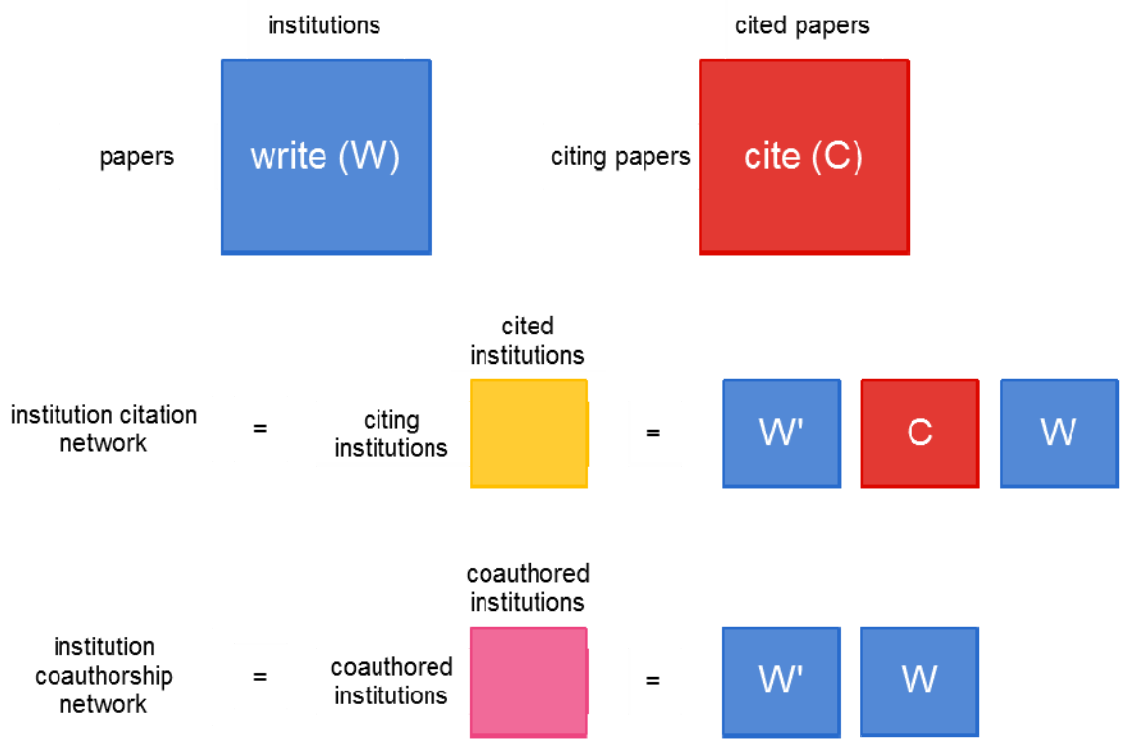


Figure 2. An illustration of the construction of citation and coauthorship networks

For articles with multiple authors and multiple affiliations, each unique affiliation pair was calculated. For example, given citing article A with three affiliations (inst_a, inst_b, and inst_c), three coauthorship relations are formed: inst_a-inst_b, inst_a-inst_c, and inst_b-inst_c. The coauthorship networks are then constructed based on these coauthor and citation relations. If A cites article B which has two affiliations (inst_b and inst_d), then the following six citation links are formed: inst_a-inst_b, inst_a-inst_d, inst_b-inst_b, inst_b-inst_d, inst_c-inst_b, and inst_c-inst_d. The citation networks are subsequently constructed based on these coauthor and citation relations. Similar to the approach of Boyack and Klavans (2010), these citation networks are then concatenated with their transposed networks, form symmetric networks.

The construction of co-citation and bibliographic coupling networks

The construction of co-citation and bibliographic coupling networks is also based on the basic networks *C* and *W*. First, two intermediate matrices are produced. A_{ij}=1 if institution *j* is cited by paper *i*, and 0 otherwise. B_{ij}=1 if paper *j* is cited by institution *i*, and 0 otherwise. Institution co-citation network is then *A'*A*, and institution bibliographic coupling network is *B*B'*. The procedure is illustrated in Figure 3.

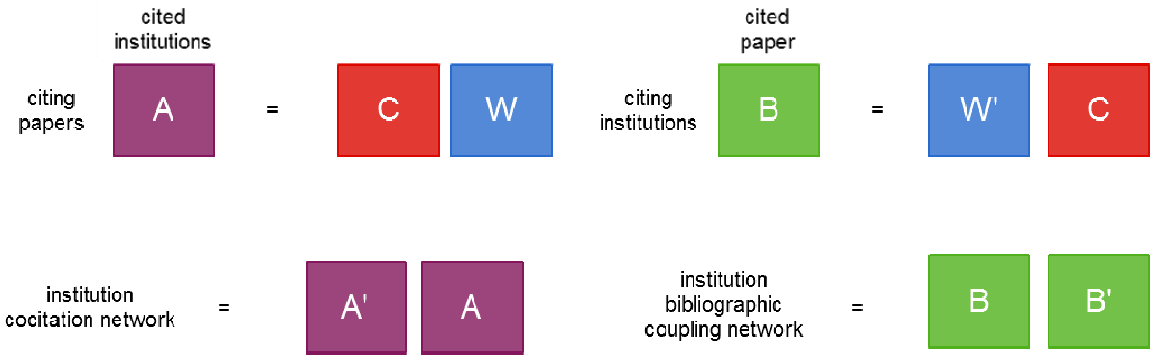


Figure 3. An illustration of the construction of co-citation and bibliographic coupling networks

The construction of co-word networks

The construction of co-word networks uses the basic matrices *W* and papers_title-words adjacency matrix *T*. In *T*, T_{ij}=1 if paper *i* contains title word *j*, and 0 otherwise[4]. D_{ij}=1 if institution *i* contains title word *j*, and 0 otherwise (note that *D* is also a binary matrix). Institution co-word network is then *D*D'*. The procedure is illustrated in Figure 4.

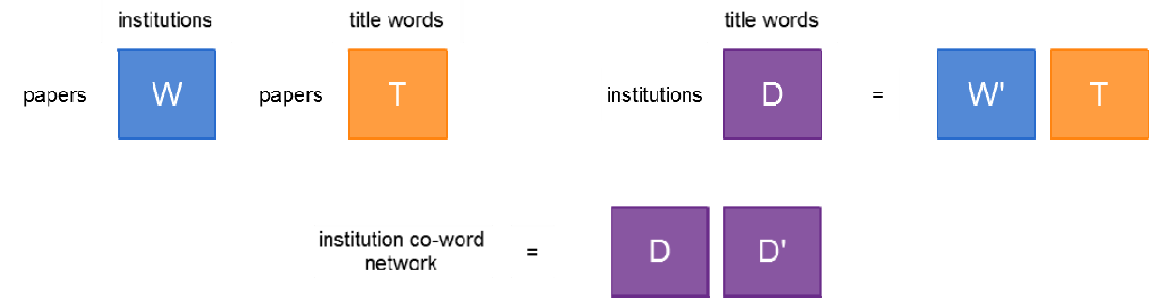


Figure 4. An illustration of the construction of co-word networks

The final treatment for coauthor networks, bibliographic coupling networks, co-citation networks, citation networks, and co-word networks is to screen out those links that are more likely to be random, as exemplified by Jarneving (2005). Random links were excluded by eliminating links whose weight is one.

The construction of topical networks

In order to quantify the topic similarity between institutions, the Author-Conference-Topic (ACT) Model (Tang, Zhang, Yao, Li, Zhang, & Su, 2008) was used. The underlying idea of the ACT Model is that if two articles share more title (or abstract) words, they have a higher probability of being on the same research topic. This can also

[4] A stop word list is used at http://ella.slis.indiana.edu/~eyan/papers/stoplist2.txt

be extended to institutions, in that if two institutions publish articles with similar title words, they are more likely to be in the same research topic. The number of topics was set at ten, and thus, each institution received a topic probability distribution: $T_i = (t_1, t_2, \ldots, t_{10})$, for institution i. A threshold is set up to replace those probabilities that fall below the average 0.1 (1/10) to 0; by doing so, the insignificant probabilities will not be counted and will thus not add noise to the similarity calculation. The topic similarity between two institutions can be calculated using cosine similarity. The cosine similarity between two institutions i and j is given by:

$$S_{ij} = \frac{T_i T_j^{'}}{\sqrt{(T_i T_i^{'})(T_j T_j^{'})}} \quad (1)$$

S_{ij} is then the edge value between institution i and institution j in the topical network. The resulting network is further dichotomized by only including those lines that have weight higher than 0.80. A simple weighting system was applied: for line values between 0.8 and 0.9, their values were set to 1 and for line values between 0.9 and 1, their values were set to 2.

Methods

Clustering and mapping technique

VOSviewer clustering technique (Waltman, Van Eck, & Noyons, 2010) is selected. It is developed based on Clauset, Newman, and Moore's (2004) algorithm for weighted networks. This algorithm incorporated *modularity*, a measurement proposed by Newman and Girvan (2004) to evaluate the community structures.

The advantage of VOSviewer clustering technique is that it unifies mapping and clustering approaches by solving the issue of minimizing:

$$V(x_1, \ldots x_n) = \sum_{i<j} S_{ij} d_{ij}^2 - \sum_{i<j} d_{ij} \quad (2)$$

where $S_{ij} = \frac{2mA_{ij}}{k_i k_j}$ (m denotes the total number of links in the network; k_i is the degree of a vertex i in a weighted network). d_{ij} has two options: for mapping, d_{ij} denotes the distance between nodes i and j in a p-dimensional map: $d_{ij} = \|x_i - x_j\| = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{ik})^2}$; for clustering, $d_{ij} = \begin{cases} 0 & \text{if } x_i = x_j \\ 1/\gamma & \text{if } x_i \neq x_j \end{cases}$ where γ is called the resolution parameter.

Waltman, Van Eck, and Noyons (2010) have shown that minimizing V is equivalent to maximizing the *modularity* for weighted networks. It can be found that *modularity* for

weighted networks is a special case when the resolution parameter γ and the weights w_{ij} are set equal to one.

Distance measurement

Cosine distance is chosen as it is a standard inquiry into the distance between two frequency vectors. Cosine distance is calculated based on cosine similarity (formula (1)). The way of calculating the cosine distance between two matrices is to transform the two matrices into two long vectors and to calculate the cosine distance using those vectors (transform the *m*-by-*m* matrix into *m*m*-by-1 vector). For example, the cosine distance between two matrices $BGcoupling_{3783*3783}$ and $Citation_{3783*3783}$ is 1-*cosine* ($BGcoupling_{14311089*1}$, $Citation_{14311089*1}$).

Results

A case study of 20 institutions

Clustering results

Network comparisons were conducted at two levels. We first chose 20 most cited institutions and observed what clusters they were grouped into and how they were visualized in a two-dimensional map. We then used cosine distance to calculate the network similarities for the whole networks.

Table 2 shows the number of links, total number of link weights, network density, number of clusters, and size of largest clusters for the six types of scholarly networks under the three time periods. The features mentioned above provide basic information for network comparisons.

Table 2. Basic network characteristics

		No. of links	Sum of link weights	Density	No. of clusters	Size of the largest cluster
1991-2000	BGcoupling	20,179	132,187	0.0048	51	285
	Citation	3,512	14,913	0.0008	108	116
	Co-citation	14,330	104,921	0.0034	48	181
	Topic	323,113	548,842	0.0765	-[5]	-
	Coauthor	302	788	0.0001	98	49
	Co-word	305,104	1,247,191	0.0723	-	-
2001-2005	BGcoupling	26,170	196,282	0.0058	44	262
	Citation	4,882	19,434	0.0011	44	136
	Co-citation	18,136	118,017	0.0040	42	216

[5] Number of clusters and size of the largest cluster were only calculated for co-word and topical networks in 2006-2010.

	Topic	356,125	609,095	0.0786	-	-
	Coauthor	530	1,451	0.0001	102	44
	Co-word	437,401	1,908,021	0.0966	-	-
2006-2010	BGcoupling	62,263	397,381	0.0087	54	518
	Citation	7,970	29,152	0.0011	58	218
	Co-citation	30,027	219,808	0.0042	44	361
	Topic	576,065	1,009,857	0.0805	25	522
	Coauthor	785	2,225	0.0001	112	105
	Co-word	835,198	3,769,316	0.1168	272	77

Co-word networks have the highest density and number of links, followed by topical networks, bibliographic coupling networks, co-citation networks, and citation networks. Coauthorship networks have the lowest density and number of links, where the density is only a tenth of the co-citation networks and a hundredth of the topical networks, suggesting that real social connections are more difficult to establish than similarity approximations. This also shows that author level collaborations were largely made within institutions, in that authors are more inclined to collaborate with other authors who are collocated, making inter-institutional collaborations less common.

Densities of the networks also affect the clustering results. The number of clusters for bibliographic coupling networks, citation networks, and co-citation networks are around 50, and there are more than 100 clusters for coauthorship networks. Since the same resolution parameter γ was chosen, the higher number of clusters indicates the existence of quite a few loosely connected and locally situated sub-networks. The results are suggestively desirable, as Dunbar (1998) predicted that 150 is roughly the upper limit of a well-functioning human community. Other studies also found that smaller communities are desirable (Allen, 2004; Leskovec, Lang, Dasgupta, & Mahoney, 2008). Leskovec et al. (2008) found that communities of size beyond 100 become less community-like, "with a roughly inverse relationship between community size and optimal community quality" (p. 1). By comparison, denser networks, such as topical networks, exhibit more generic characteristics and thus yield fewer clusters.

Table 3 lists the top 20 institutions based on the number of citations during 2006-2010 (a detailed calculation can be found in Yan and Sugimoto, 2011). The numbers in the third to sixth columns are cluster IDs of each institution calculated through VOSviwer clustering technique (co-word and topical networks were not visualized due to their densities).

Table 3. Clustering results of top institutions

Idx	Institution name	BGcoupling	Citation	Co-citation	Coauthor
1	GEORGIA STATE UNIV,ATLANTA	4	29	4	20

2	HUNGARIAN ACAD SCI,HUNGARY	1	1	2	60
3	UNIV GEORGIA,ATHENS	9	42	4	88
4	UNIV MINNESOTA,MINNEAPOLIS	4	48	4	79
5	UNIV WESTERN ONTARIO,CANADA	15	6	1	85
6	INDIANA UNIV,BLOOMINGTON	7	15	33	94
7	FLORIDA STATE UNIV,TALLAHASSEE	7	30	28	47
8	UNIV BRITISH COLUMBIA,CANADA	5	16	4	104
9	UNIV OKLAHOMA,NORMAN	20	11	30	13
10	UNIV SHEFFIELD,ENGLAND	2	13	26	11
11	UNIV MARYLAND,COLLEGE PK	7	3	28	75
12	UNIV MICHIGAN,ANN ARBOR	9	28	27	37
13	DREXEL UNIV,PHILADELPHIA	14	42	13	35
14	KATHOLIEKE UNIV LEUVEN,BELGIUM	1	1	2	60
15	UNIV S FLORIDA,TAMPA	7	43	9	57
16	ROYAL SCH LIB & INF SCI,DENMARK	1	9	9	78
17	LEIDEN UNIV,NETHERLANDS	1	1	2	14
18	UNIV ARIZONA,TUCSON	18	32	1	97
19	UNIV PITTSBURGH,PITTSBURGH	12	5	3	79
20	UNIV ILLINOIS,URBANA	2	6	1	70

In the bibliographic coupling network, institutions such as HUNGARIAN ACAD SCI,HUNGARY, ROYAL SCH LIB & INF SCI,DENMARK, KATHOLIEKE UNIV LEUVEN,BELGIUM, and LEIDEN UNIV,NETHERLANDS are more likely to cite the same articles (cluster ID: 1); in the citation network, institutions such as UNIV WESTERN ONTARIO,CANADA and UNIV ILLINOIS,URBANA are more likely to cite each other (cluster ID: 6); in the co-citation network, institutions such as GEORGIA STATE UNIV,ATLANTA, UNIV GEORGIA,ATHENS, UNIV BRITISH COLUMBIA,CANADA, and UNIV MINNESOTA,MINNEAPOLIS are more likely to be cited by the same articles (cluster ID: 4); in the coauthorship network, institutions such as UNIV MINNESOTA,MINNEAPOLIS and UNIV PITTSBURGH,PITTSBURGH maintain tighter collaboration relationships (cluster ID: 79).

Mapping results

The VOSviewer mapping technique is a weighted version of multidimensional scaling (Van Eck et al., 2010). The following four figures (Figure 5 to Figure 8)[6] are used to understand how different scholarly networks would yield different mapping results about the 20 institutions. Numbers in circles are institutions' index number (first column in Table 3).

[6] Interactive visualizations can be found at http://info.slis.indiana.edu/~eyan/papers/citation/

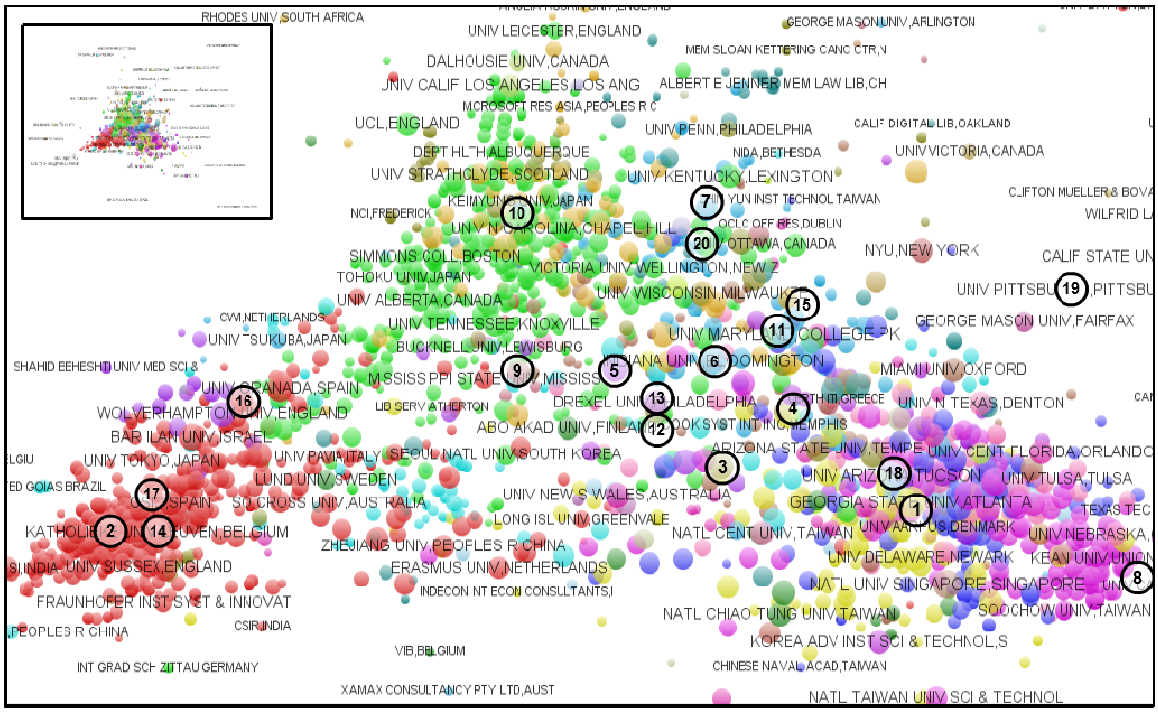


Figure 5. A visualization of the bibliographic coupling network (2006-2010)

In the bibliographic coupling network (Figure 5), three bibliometric institutions (e.g., 2, 14, and 17) are located at the left side of the map. Information system institutions (e.g., 1, 8, and 18) are located at the right side of the map. Institutions specializing in other topics in LIS, such as information retrieval (e.g., 13 and 12) and library science (e.g., 7 and 20), are found in the middle of the map.

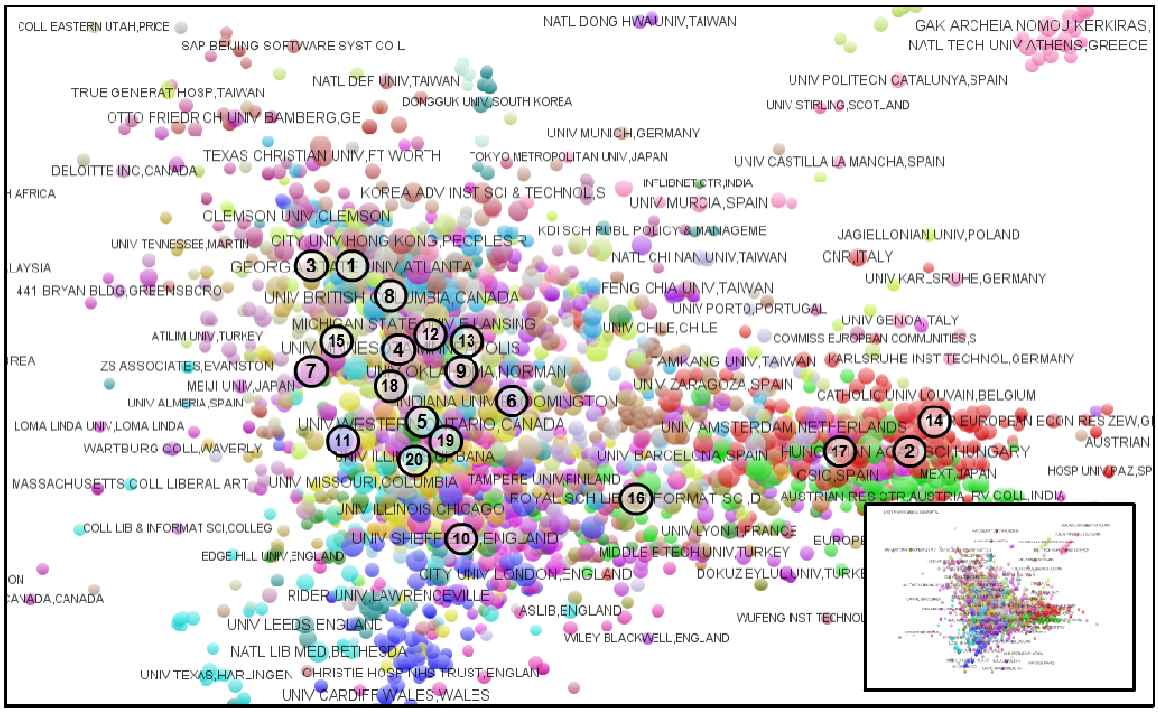


Figure 6. A visualization of the citation network (2006-2010)

In the citation network (Figure 6), the three bibliometric institutions are still collocated, but are now found at the right side of the map. Information system institutions (e.g., 1, 3, and 8) are located at the left side of the map. Similar to Figure 5, ROYAL SCH LIB & INF SCI,DENMARK (16) is located between bibliometric institutions and information system institutions due to its multiple focuses on bibliometrics, information retrieval, and library science topics.

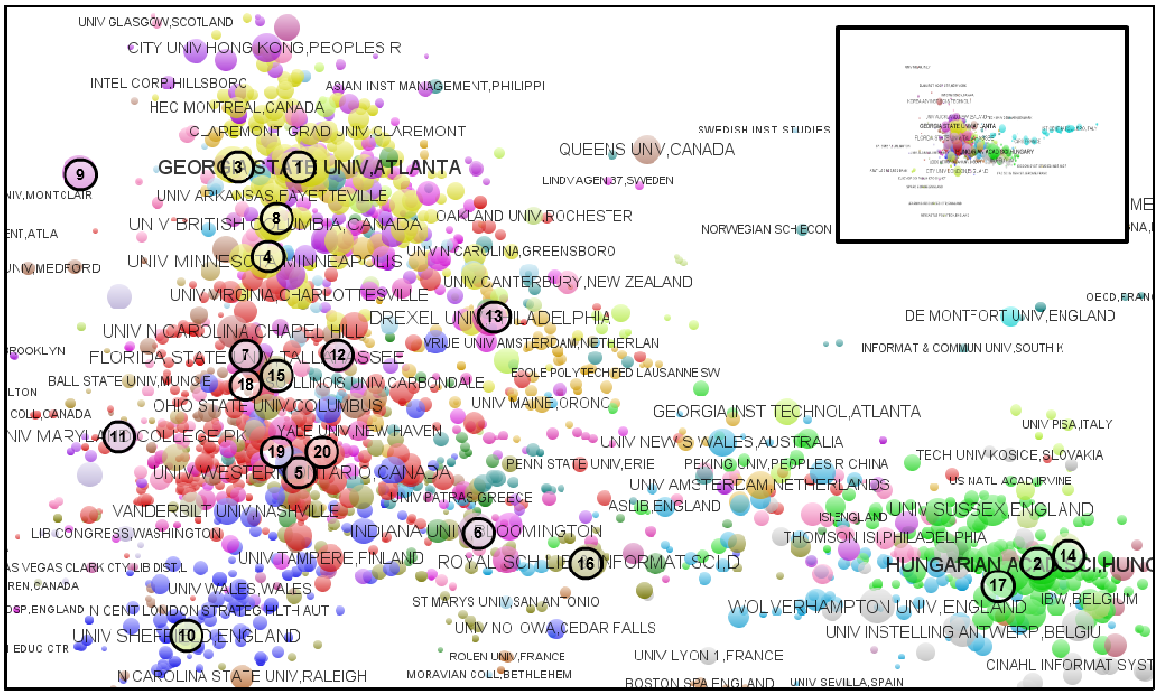


Figure 7. A visualization of the co-citation network (2006-2010)

In the co-citation network (Figure 7), the division between bibliometric institutions and other institutions becomes more evident. Bibliometric institutions (e.g., 2, 14, and 17) are located at the far right side of the map. Information system institutions are located at the upper left of the map, and library science institutions are located at the lower left of the map. UNIV SHEFFIELD,ENGLAND (10) is not collocated on the map with other top institutions. By reading its nearby institutions, we can find that the majority are British institutions. Geographical location can thus be a factor in institutional citation behaviors (Yan & Sugimoto, 2011), in that British institutions, in this case, are more likely to cite, co-cite, or be co-cited by other British institutions.

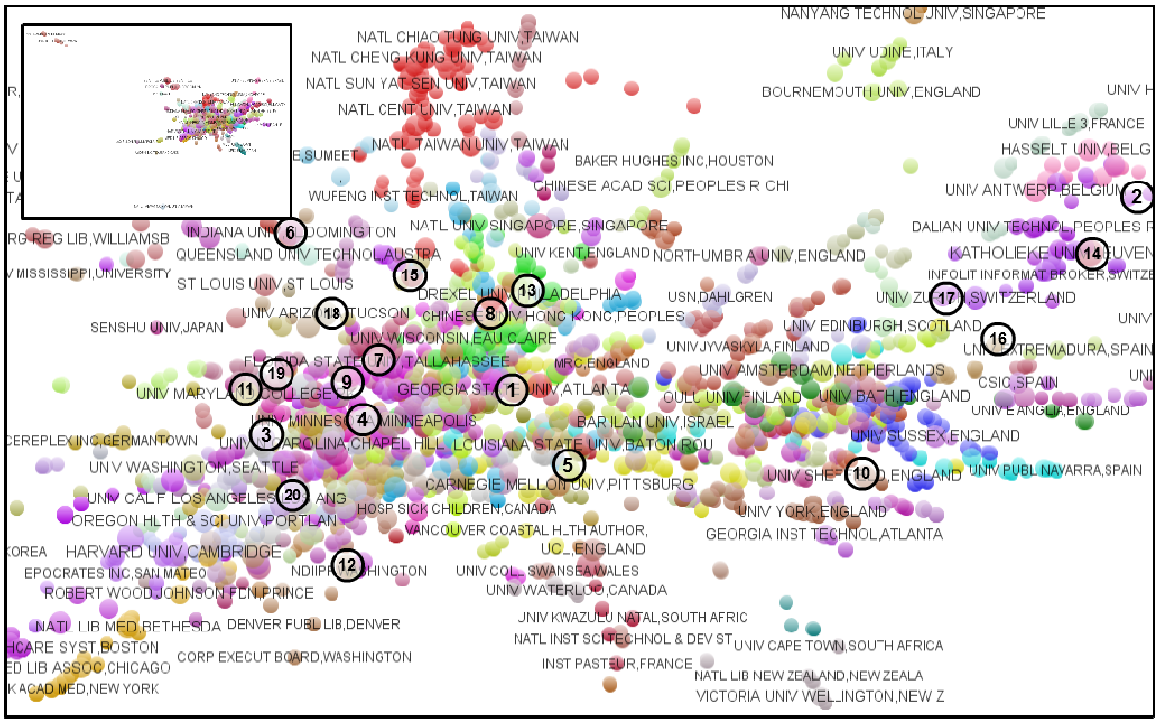


Figure 8. A visualization of the coauthorship network (2006-2010)

In the coauthorship network (Figure 8), more clusters can be found but the general locations of top institutions resemble the previous maps, in that bibliometric institutions are located at one side of the map and information system and library science institutions are located at the other side. It can be inferred that citation and collaboration relations are interweaving, meaning that if two institutions have collaboration relations (or citation relations), they are more likely to cite (or collaborate with) one another than in the absence of such relations.

Network comparisons of all institutions

Table 4 shows the network similarities calculated based on cosine distance. Cosine distances range from 0 to 1, where a value of zero is an indication of two identical networks and a value of one is an indication of two totally dissimilar networks.

Table 4. Network similarities measured by cosine distance

	BGcoupling	Citation	Co-citation	Topic	Coauthor	Co-word
BGcoupling	-	0.17[7]	0.28	0.97	0.93	0.45
		0.34	0.56	0.99	0.90	0.79
		0.29	0.42	0.97	0.76	0.50
Citation	0.17	-	0.01	0.99	0.96	0.59

[7] The cosine distance between BGcoupling and Citation networks in 1991-2000; 0.34 is the cosine distance for the two networks in 2001-2005 and 0.29 is the cosine distance for the two networks in 2006-2010.

	0.34		0.19	0.98	0.84	0.61
	0.29		0.26	0.98	0.77	0.59
Co-citation	0.28	0.01	-	0.99	0.99	0.65
	0.56	0.19		0.97	0.92	0.60
	0.42	0.26		0.97	0.87	0.59
Topic	0.97	0.99	0.99	-	0.99	0.93
	0.99	0.98	0.97		0.99	0.94
	0.97	0.98	0.97		0.99	0.94
Coauthor	0.93	0.96	0.99	0.99	-	0.91
	0.90	0.84	0.92	0.99		0.87
	0.76	0.77	0.87	0.99		0.88
Co-word	0.45	0.59	0.65	0.93	0.91	-
	0.79	0.61	0.60	0.94	0.87	
	0.50	0.59	0.59	0.94	0.88	

Based on cosine distances, bibliographic coupling networks and citation networks have the smallest value and thus have the highest similarity; topical networks and coauthorship networks have the highest value and thus have the lowest similarity. A finding can therefore be made that topical networks and coauthorship networks set two boundaries for all six networks. Collaboration and topical adjacencies are different scholarly communication channel, where collaborations are social interactions, and topical adjacency is a form of cognitive approximation derived from knowledge recognition and identification.

Discussion

In this discussion section, we first validate the use of scholarly networks in studies of scholarly communications and science policy making. Second, since each scholarly network has its own applications, it is necessary to use appropriate distinctions to uncover their properties. Multidimensional scaling technique is then used to examine how different scholarly networks relate to each other. A recommendation is finally made that hybrid networks be developed as they are capable of capturing varied aspects of research interactions.

The use of scholarly networks in studies of scholarly communication and science policy making

Before network theories were introduced to bibliometrics, accumulative citation counting was widely used in scientific evaluation. In the same vein of research, several citation-based indicators were proposed, such as Journal Impact Factor (Hirst, 1978) and *h*-index (Hirsch, 2005). The accumulative citation counting and citation-based indicators equated all citations to have the same weight, without consideration of the citing papers, citing authors, or citing journals. But this equal counting mechanism has been questioned, where scholars (e.g., Pinski & Narin, 1976; Cronin, 1984; Bollen, Rodriguez, & Van de

Sompel, 2006; Yan, Ding, & Sugimoto, 2011) have argued that it is more reasonable to differentiate citation weights based on the source of endorsement. This tension has largely been alleviated by the construction of different types of scholarly networks and the invention of various network-based bibliometric indicators. Compared to traditional citation counting, scholarly networks have the advantage of considering the source of the citation endorsement. In this way, scholarly networks can capture the complex research communication and interaction more precisely.

In addition to scientific evaluation, scholarly networks contribute to other realms of scholarly communication and science policy making. For instance, coauthorship networks provide an accurate and expedite medium, allowing scientists and scholars to explore various intriguing questions pertinent to scientific collaboration and research communities (e.g., Logan & Shaw, 1991; Luukkonen, Persson, & Sivertsen, 1992; Newman, 2004; Moody, 2004; Ahn, Bagrow, & Lehmann, 2010); co-citation networks, bibliographic coupling networks, and co-word networks have been used to identify research specialties, examine interdisciplinarities, and map the backbone of science (e.g., Kessler, 1963; Small, 1973; White & Griffith, 1981; White & McCain, 1998; Boyack, Klavans, & Börner, 2005; Chen, 2006); and citation networks have been used to study knowledge flows and find knowledge paths in science (e.g., Jaffe, Trajtenberg, & Henderson, 1993; Narin, Hamilton, & Olivastro, 1997; Rinia, Van Leeuwen, & Bruins, 2001; Chen & Hicks, 2004; Mehta, Rysman, & Simcoe, 2010; Yan & Sugimoto, 2011).

Distinctions in scholarly networks

In an important review article on networks, Newman (2003) distinguished four categories for real-world networks: social networks (e.g., collaboration networks), information networks (e.g., citation networks), technical networks (e.g., Internet router networks), and biological networks (e.g., protein networks). Based on such divisions, two types of scholarly networks can be distinguished, social networks vs. information networks. In social networks such as coauthorship networks, a node is a social actor (i.e., an author), yet in information networks, a node is usually an artifact, such as a paper, a journal, or an institution. In addition to "social networks vs. information networks", another distinction of "real connection-based networks vs. artificial connection-based networks" can be made. Coauthorship networks and citation networks are constructed based on real connections, whereas co-citation, bibliographic coupling, topical, and co-word networks are constructed based on artificial connections[8], usually in the form of similarity measurements. These scholarly networks can also be viewed from their edge types: collaboration-based, citation-based, or word-based. Citation-based scholarly networks include citation networks, co-citation networks, and bibliographic coupling networks;

[8] Even though in author co-citation/BGcoupling networks a node can be an author, such a node is considered as an aggregator of papers (see how the networks were constructed in the Method section).

word-based scholarly networks include topical networks and co-word networks; and collaboration-based networks include coauthorship networks. These distinctions (citation-based networks vs. non-citation-based networks; social networks vs. information networks, real connection-based networks vs. artificial connection-based networks, as shown in Figure 9) are helpful in understanding how different types of scholarly networks relate to each other.

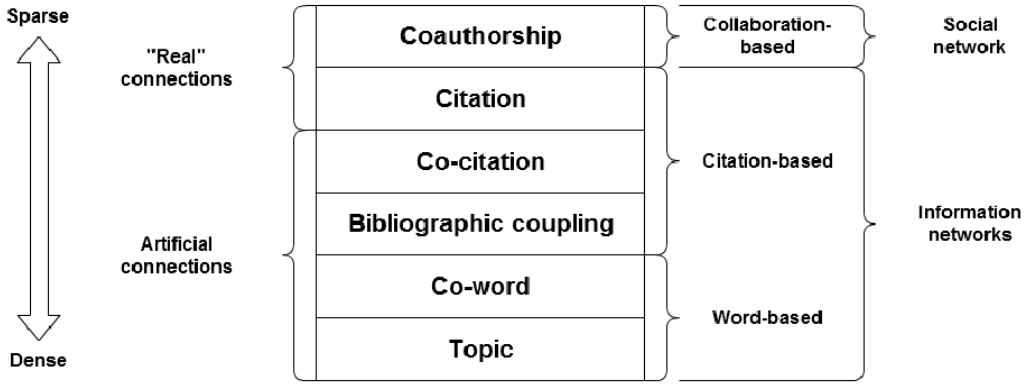


Figure 9. Viewing scholarly networks from different perspectives

Scholarly networks similarities

Table 5 shows the similarity rankings for each pair of networks based on cosine distance in the 2006-2010 period. For example, the second row in Table 5 can be read as: bibliographic coupling networks are most similar to citation networks, followed by co-citation networks, co-word networks, coauthorship networks, and topical networks. The second column can be read as for coauthor and co-word networks, bibliographic coupling networks are most similar to them, and for citation, co-citation, and topical networks, bibliographic coupling networks are the second most similar to them.

Table 5. Ranking of network similarities (calculations based on Cosine Distance)

	BGcoupling	Citation	Co-citation	Topic	Coauthor	Co-word
BGcoupling	-	1	2	5	4	3
Citation	2	-	1	5	4	3
Co-citation	2	1	-	5	4	3
Topic	2	4	2	-	5	1
Coauthor	1	2	3	5	-	4
Co-word	1	2	2	5	4	-

Topical networks and coauthorship networks are found here to have the lowest similarity; co-citation networks and citation networks have low similarities with coauthorship networks; co-citation networks and citation networks have high similarity; bibliographic coupling networks and co-citation networks have high similarity; co-word networks and

topical networks have high similarity, and so forth. We use multidimensional scaling (MDS) to generalize the findings in Table 5. For MDS, the input data are cosine distances between networks in 2006-2010.

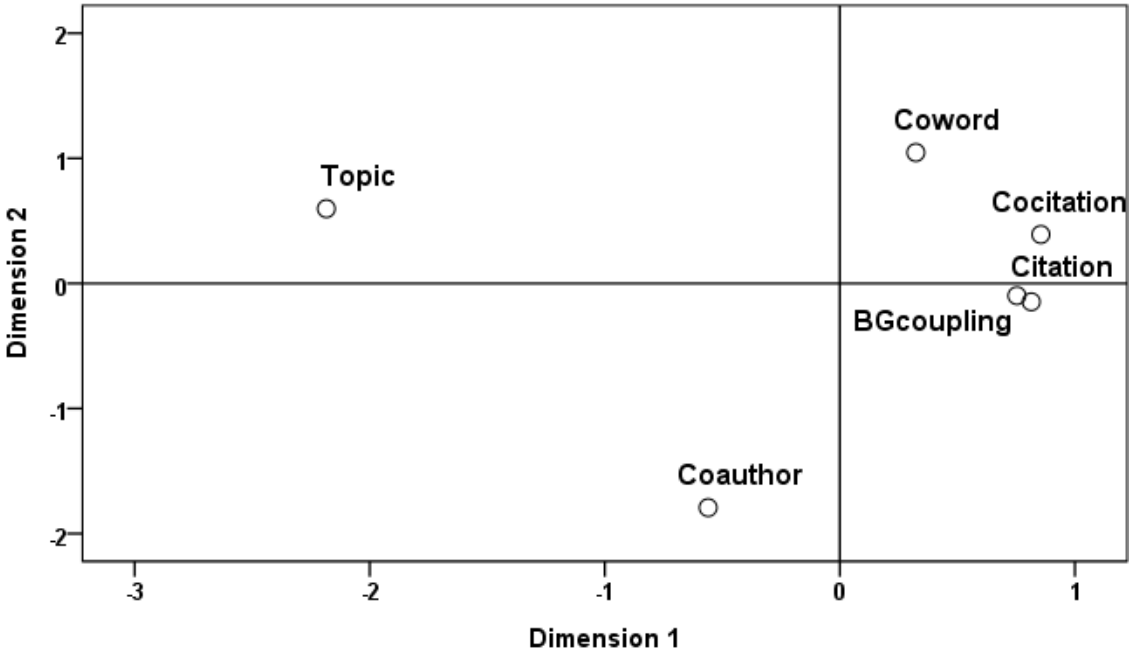


Figure 10. A visualization of network similarities as exemplified by MDS (stress value: 0.18)

The two dimensions are responsible for 84 percent of total variances. Dimension 1 can be interpreted as "network remoteness". Topical networks are remote from all other networks. Each column in Table 5 shows how remote each network is to other networks, where topical networks are the most dissimilar to the remaining five networks, followed by co-word networks and coauthorship networks. The remoteness of the topical network cannot be fully attributed to its density, as the co-word network is also a dense network but it is not as remote as the topical network. On the other hand, BGcoupling networks, co-citation networks, and citation networks are closer to other networks. Dimension 1 can also be perceived as "non-citation-based vs. citation-based" because topical networks, co-word networks, and coauthorship networks are non-citation-based, and BGcoupling networks, co-citation networks, and citation networks are citation-based. Dimension 2 can be interpreted as "social vs. cognitive" wherein "social" refers to social connections such as collaboration relations, and "cognitive" mainly refers to similarity of lexical semantics. For example, co-word and co-citation networks have been used to identify research fields, map the backbone of science, or portray intellectual landscapes.

Hybrid scholarly networks

We recommend that in order to capture varied aspects of research interactions, different types of scholarly networks need to be combined to form a hybrid network. Intuitively, three hybrid scholarly networks can be constructed. The first hybrid scholarly network focuses on Dimension 1 "non-citation-based vs. citation-based" in Figure 10. Two scholarly networks, topical networks and co-citation networks can be linearly combined, thus incorporate the largest variance:

$$H_{topical+co-citation} = \alpha M_{topical} + (1-\alpha) M_{co-citation} \quad (3)$$

The second hybrid scholarly network focuses on Dimension 2 "social vs. cognitive" where two scholarly networks, coauthorship networks and co-word networks, can be linearly combined:

$$H_{coauthor+co-word} = \alpha M_{coauthor} + (1-\alpha) M_{co-word} \quad (4)$$

The third hybrid scholarly networks focuses on both dimensions, and thus integrates all six scholarly networks, where $a+b+c+d+e+f=1$:

$$H = aM_{coauthor} + bM_{co-word} + cM_{BGcoupling} + dM_{citation} + eM_{co-citation} + fM_{topical} \quad (5)$$

In this hybrid combination, when a=b=c=d=0, f=α , and e=1-α , equation (5) will become (3); when c=d=e=f=0, a=α , and b=1-α , equation (5) will become (4). Equation (5), therefore, is able to linearly combine different types of scholarly networks in a flexible way.

Janssens, Glänzel, and De Moor (2008), however, argued that the weighted linear combinations may "neglect different distributional characteristics of various data sources" (p. 612). Therefore, appropriate thresholding and/or dichotomization is required for networks of diverse densities.

In addition to linear combinations, scholars have proposed other approaches to form hybrid scholarly networks. For example, Cao and Gao (2005) used a feature selection method to classify scientific papers. Their method consists of two steps: it first applied a content-based classification, and then iteratively updated the labeling of unknown documents using papers' cited references. Janssens, Glänzel, and De Moor (2008) integrated a term-by-document matrix and a cited_references-by-document matrix through Fisher's inverse chi-square method. Liu et al. (2010) presented a framework of hybrid clustering to combine lexical and citation data for journal sets analysis. Two approaches, clustering ensemble and kernel-fusion clustering, were utilized in their framework. Boyack and Klavans (2010) developed a bibliographic coupling-based citation-text hybrid approach that couples both references and words from titles/abstracts.

Yan, Ding, and Jacob (2012) coupled two paper-to-paper matrices, where in the paper-to-paper (author) matrix, a cell value denotes the number of shared authors, and in the paper-to-paper (word) matrix, a cell value denotes the number of shared title words. Through matching the two matrices, the mutual dependency of research topics and research communities in library and information science was uncovered.

Conclusion

Previous studies on scholarly networks usually chose one type of network at one aggregation level. But the choice of networks types can be inconsistent, and the findings have been discrete, and cannot therefore be generalized to address a wider spectrum of research questions. This study provides a solution to this problem by exploring the similarity among six types of scholarly networks aggregated at the institution level.

We find that topical networks and coauthorship networks have the lowest similarity, and these two types of networks set two boundaries (social and cognitive) for all six networks; co-citation networks and citation networks have high similarity; bibliographic coupling networks and co-citation networks have high similarity; co-word networks and topical networks have high similarity, and so forth. Factors that contribute to the similarities are edge types ("real" connections vs. artificial connections; citation-based connections vs. non-citation-based connections) and network types (social networks vs. information networks). In addition, through MDS, two dimensions can be identified and used to describe the six types of scholarly networks, where Dimension 1 can be interpreted as "non-citation-based vs. citation-based", and Dimension 2 can be interpreted as "social vs. cognitive".

A recommendation is made herein that hybrid scholarly networks can more comprehensively capture the complex research communication and interaction. The findings of this study indicate that future research on this topic would benefit from evaluating different approaches to hybrid networks or heterogeneous networks through the possible application of "golden standards" (such as award lists or expert judgments) that can help determine which approach yields more precise clustering results and useful information for scientific evaluations and science policy making.

Acknowledgements

The authors would like to thank Dr. Stanley Wasserman for his guidance in this research project. The authors would also like to thank Ludo Waltman of CWTS, Leiden University for his comments on methods used in this article.

References

Allen, C. (2004). Life with alacrity: The Dunbar number as a limit to group sizes. Retrieved January 23, 2012 from http://www.lifewithalacrity.com/2004/03/the_dunbar_numb.html

Bollen, J., Rodriguez, M. A., & Van De Sompel, H. (2006). Journal status. *Scientometrics, 69*(3), 669-687.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology, 61*(12), 2389-2404.

Boyack, K. W., Klavans, A. R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics, 64*(3), 351-374.

Cao, M., & Gao, X. (2005). Combining contents and citations for scientific document classification. *Advances in artificial intelligence, 3809*, 143-152. Retrieved January 23, 2012 from http://www.springerlink.com/index/p81585434700r161.pdf

Chen, C. M. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359-377.

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics, 1*(1), 8-15.

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large network. *Physical Review E, 70*(6), 066111.

Ding, Y., & Cronin, B. (2011). Popular and/or prestigious? Measures of scholarly esteem. *Information Processing and Management, 47*(1), 80-96.

Ding, Y., Chowdhury, G., & Foo, S. (2000). Journal as markers of intellectual space: Journal co-citation analysis of information retrieval area, 1987-1997. *Scientometrics, 47*(1), 55-73.

Dunbar, R. (1998). *Grooming, gossip, and the evolution of language*. Boston, MA: Harvard University Press.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569-16572.

Hirst, G. (1978). Discipline impact factors: Method for determining core journal lists. *Journal of the American Society for Information Science, 29*(4), 171-172.

Jaffe, A. B., Trajtenberg, M., & Henderson, A. D. (1993). Geographical localization of knowledge spillovers by patent citations. *Quarterly Journal of Economics, 108*(3), 577-599.

Janssens, F., Glänzel, W., & De Moor, B. (2008). A hybrid mapping of information science. *Scientometrics, 75*(3), 607-631.

Jarneving, B. (2005). A comparison of two bibliometric methods for mapping of the research front. *Scientometrics, 65*(2), 245-263.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*(1), 10-25.

Leskovec, J., Lang, K. J., Dasgupta, A., & Mahoney, M. W. (2008). Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. Retrieved January 23, 2012 from http://arxiv.org/abs/0810.1355

Liu, X., Yu, S., Janssens, F., Glanzel, W., Moreau, Y., & De Moor, B. (2010). Weighted hybrid clustering by combining text mining and bibliometrics on a large-scale journal database. *Journal of the American Society for Information Science and Technology, 61*(6), 1105-1119.

Logan, E. L., & Shaw, W. M. (1991). A bibliometric analysis of collaboration in a medical specialty. *Scientometrics, 20*(3), 417-426.

Luukkonen, T. (1997). Why has Latour's theory of citations been ignored by the bibliometric community? Discussion of sociological interpretations of citation analysis. *Scientometrics, 38*(1), 27-37.

Ma, N., Guan, J., & Zhao, Y. (2008). Bringing PageRank to the citation analysis. *Information Processing and Management, 44*(2), 800-810.

Mehta, A., Rysman, M., & Simcoe, T. (2006). Identifying the age profile of patent citations. *Social Science Research Network, 25*(7), 1179-1204.

Milojević, S., Sugimoto, C. R., Yan, E., & Ding, Y. (2011). The cognitive structure of library and information science: Analysis of article title words. *Journal of the American Society for Information Science and Technology, 62*(10), 1933-1953.

Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review, 69*(2), 213-238.

Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between U.S. technology and public science. *Research Policy, 26*(3), 317-330.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review, 45*(2), 167-256.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America, 101*(suppl. 1), 5200-5205.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*(2), 026113.

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management, 12*(5), 297-312.

Radicchi, F., Fortunato, S., Markines, B., & Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E, 80*(5), 056103.

Rinia, E. J., Van Leeuwen, T. N., & Bruins, E. E. W. (2001). Citation delay in interdisciplinary knowledge exchange. *Scientometrics, 51*(1), 293-309.

Sayyadi, H., & Getoor, L. (2009). FutureRank: Ranking scientific articles by predicting their future PageRank. In *Proceedings of the Ninth SIAM International Conference on Data Mining*. Retrieved January 23, 2012 from http://www.siam.org/proceedings/datamining/2009/dm09_050_sayyadih.pdf

Small, H. G. (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*(4), 265-269.

Sun, Y., Barber, R., Gupta, M., Aggarwal, C., & Han, J. (2011). Co-author relationship prediction in heterogeneous bibliographic networks. In *Proceedings of 2011 International Conference on Advances in Social Network Analysis and Mining*, July 35-27, 2011, Kaohsiung, Taiwan.

Sun, Y., Yu, Y., & Han, J. (2009). Ranking-based clustering of heterogeneous information networks with star network schema. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. June 28-July 1, 2009, Paris, France.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and mining of academic social networks. In *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp.990-998). August 24-27, Las Vegas, NV.

Van Eck, N. J., Waltman, L., Dekker, R., & Van den Berg, J. (2010). A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS. *Journal of the American Society for Information Science and Technology, 61*(12), 2405-2416.

Walker, D., Xie, H., Yan, K. K., & Maslov, S. (2007). Ranking scientific publications using a simple model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, P06010, doi:10.1088/1742-5468/2007/06/P06010

Waltman, L., Van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics, 4*(4), 629-635.

Waltman, L., Yan, E., & Van Eck., N. J. (2011). A recursive field-normalized bibliometric performance indicator: An application to the field of library and information science. *Scientometrics, 89*(1), 301-314.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, UK: Cambridge University Press.

West, J. D., Bergstrom, T. C., & Bergstrom, C. T. (2010). The Eigenfactor Metrics: A network approach to assessing scholarly journals. *College and Research Libraries, 71*(3), 236-244.

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science, 32*(3), 163-171.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science, 49*(4), 327-355.

Yan, E., & Ding, Y. (2011). Discovering author impact: A PageRank perspective. *Information Processing and Management, 47*(1), 125-134.

Yan, E., & Sugimoto, C. R. (2011). Institutional interactions: Exploring the social, cognitive, and geographic relationships between institutions as demonstrated through citation networks. *Journal of the American Society for Information Science and Technology, 62*(8), 1498-1514.

Yan, E., Ding, Y., & Jacob, E. K. (2012). Overlaying communities and topics: An analysis on publication networks. *Scientometrics, 90*(2), 499-513.

Yan, E., Ding, Y., & Sugimoto, C. R. (2011). P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology, 62*(3), 467-477.

Zhou, D., Orshanskiy, S. A., Zha, H., & Giles, C. L. (2007). Co-ranking authors and documents in a heterogeneous network. In *Proceedings of the Seventh IEEE International Conference on Data Mining* (pp.739-744). October 28-31, Omaha, Nebraska.

Zitt, M., Lelu, A., & Bassecoulard, E. (2011). Hybrid citation-word representations in science mapping: Portolan charts of research fields. *Journal of the American Society for Information Science and Technology, 62*(1), 19-39.