Using path-based approaches to examine the dynamic structure of discipline-level citation networks: 1997-2011

Erjia Yan[1]

College of Computing and Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, U.S.A. Phone: (215)895-1459. Fax: (215) 895-2494. Email: erjia.yan@drexel.edu

Qi Yu

School of Management, Shanxi Medical University, Taiyuan, China

Abstract

The objective of this paper is to identify the dynamic structure of several time-dependent discipline-level citation networks through a path-based method. A network data set is prepared that comprises 27 subjects and their citations aggregated from more than 27,000 journals and proceedings indexed in the Scopus database. A maximum spanning tree method is employed to extract paths in the weighted, directed, and cyclic networks. This paper finds that subjects such as Medicine, Biochemistry, Chemistry, Materials Science, Physics, and Social Sciences are the ones with multiple branches in the spanning tree. This paper also finds that most paths connect science, technology, engineering, and mathematics (STEM) fields; two critical paths connecting STEM and non-STEM fields are the one from Mathematics to Decision Sciences and the one from Medicine to Social Sciences.

Introduction

The creation and diffusion of knowledge is not dependent on a single unit. Knowledge is diffused, circulated, and utilized among various social sectors and individuals (Yan, 2014). In the past few decades, geographical, institutional, and disciplinary barriers that once confined knowledge flows are increasingly becoming permeable. Scholars demonstrated that our society has transformed from a production-based one to a knowledge-based one—or the so called knowledge societies (e.g., Knorr-Cetina, 1999) where knowledge has become the "heart of economic growth" (David & Foray, 2002, p. 9). It is thus of great importance to reexamine and understand patterns of knowledge flows in this new era.

Prior quantitative studies of knowledge have largely employed instruments such as citations, co-citations, and coauthorship relations. Perhaps, the groundbreaking work was Jaffe, Trajtenberg and Henderson's (1993) investigation of knowledge spillovers through patent citations—in particular the proximity theory was introduced to interpret the distribution of patent citations. Studies have found that physical, institutional, cognitive, technological, and geographic proximities are all associated with the knowledge diffusion in organizations (Jaffe & Trajtenberg, 1999; MacGarvie, 2005; Autant-Bernard, Mairesse, & Massard, 2007), among which geographic proximity serves as a leading factor followed by institutional proximity, as noted by Ponds, Frenken and Van Oort (2007). Cognitive proximity was examined through citation and co-keyword relations (Buter, Noyons, & Van Raan, 2010) in that both types of relations have accurately detected converging research areas in science. In terms of paper citations, Lewison, Rippon,

[1]Corresponding author

and Wooding (2005) found that citing papers in the biomedical field are progressively becoming more international and the fine line between basic and clinical research is diminishing. Using a more extensive data set that comprised all major scientific fields, Wagner and Leydesdorff (2005) revealed that the international coauthorship network is a self-organizing network guided by the principle of preferential attachment.

In addition to the proximity theory, another influential approach to understand knowledge diffusion is the epidemic models. These models can be seen as a particular case of population dynamics models (Vitanov & Ausloos, 2012). They were used to study the transmission of infectious diseases and have been adopted to study the dissemination of knowledge. Actors in the models can be represented by one of the following classes: susceptible (S), exposed (E), infected (I), skeptical (Z), and recovered (R). It is found that the SEIZ combination has the best fit with empirical data for a case study on the Feynman diagram (Hethcote, 2000). Another case study using the SI and SEI combinations found that the diffusion of publications on kinesin research was more likely to occur between disciplines with existing knowledge flows (Kiss et al., 2010). Other statistical methods that model knowledge diffusion, as noted by Geroski (2000), include the probit model which posits that the knowledge adoption rate is dependent with types of organizations and the ecologic model which is predicated upon the "twin forces of legitimation and competition" (p. 603).

Apart from these efforts, qualitative studies have been pursued to capture more nuanced social processes that shape knowledge transfer, particularly in relation to human capital—collaboration (e.g., Singh, 2005; Wuchty, Jones, & Uzzi, 2007), the mobility of employees (e.g., Almeida & Kogut, 1999; Cohen, Nelson & Walsh, 2002), and the formation of knowledge-based communities (e.g., David & Foray, 2002) have profoundly changed the production and diffusion of knowledge. For instance, Cohen, helson, and Walsh (2002) confirmed several knowledge transfer channels including conference contacts, mobility of researchers, mentorship and collaboration relationships, and virtual scholarly communication channels (e.g., publications and patens). Moreover, barriers to knowledge transfer were discovered (Szulanski, 1996; Almeida & Kogut, 1999) in that issues such as insufficient absorptive capacity, ambiguity, and inter-personal tensions are all contributing factors.

In regards to scientific knowledge, research fields are typically chosen as the unit of analysis. One thread of research has primarily focused on using co-occurrence data to map scientific fields and portray academic landscape, such as the consensus map (Klavans & Boyack, 2009), map of science (Börner et al., 2012), the overlay maps (Rafols, Porter, & Leydesdorff, 2010; Chen & Leydesdorff, 2014), and the global science map (Leydesdorff & Rafols, 2009). Yet, these co-occurrence-based maps are not effective in depicting knowledge diffusion patterns because these patterns relate to direct relations in nature. Another thread has used citations as the proxy to examine knowledge diffusion, with the notion that knowledge flows from the cited entity to the citing entity. In this vein of research, prior work used paper and patent citation data and explored knowledge flows across geographic locations (e.g., Jaffe, Trajtenberg, & Henderson, 1993), across disciplines (Van Leeuwen & Tijssen, 2000; Yan, Ding, Cronin, & Leydesdorff, 2013; Yan, 2014), among industries and organizations (e.g., Bacchiocchi & Montobbio, 2007), among academic institutions (Börner, Penumarthy, Meiss, & Ke, 2006), and between papers and patents (Narin, Hamilton, & Olivastro, 1997; Chen & Hicks, 2004).

An effective way to interpret these citation-based studies is through the trading metaphor. It makes analogies to international trade in that a discipline is a trading unit and can export knowledge via

incoming citations and import knowledge via outgoing citations (Yan et al., 2013). Using this metaphor, patterns of knowledge flows in information and library science (Cronin & Davenport, 1989; Cronin & Pearson, 1990; Larivière, Sugimoto & Cronin, 2012), statistics and probability (Stigler, 1994), management science (Lockett & McWilliams, 2005), and all disciplines in sciences and social sciences (Yan et al., 2013; Yan, 2014) were examined. These studies found that while some disciplines tended to become knowledge exporters and enjoyed knowledge surplus, other disciplines tended to become knowledge importers and experienced knowledge deficit (e.g., Yan et al., 2013).

By surveying these studies, it is found that the focus was primarily on the actors but not on knowledge paths. Furthermore, most studies pertained to case studies of specified topics and an aggregated overview of disciplinary knowledge flows is lacking from the literature. To fill the gap, this study employs the maximum spanning tree (MST) approach to identify discipline-level knowledge paths in weighted, directed, and cyclic networks. This approach, when applied to a 15-year citation data set awarded by Elsevier, identifies the dynamic patterns of paths in the spanning trees. Results from this study will provide quantitative evidences to support succeeding studies on disciplinary knowledge flows.

Data and methods

Data

The data set was awarded by the Elsevier Bibliometrics Research Program[2]. The intermediary data file was a journal-to-journal citation matrix for all journals and proceedings indexed in the Scopus database with a two-year citation window; that is, citations in year *t* to papers published in year *t*-2. Data on the following cited/citing years were therefore obtained: 1997/1999, 2000/2002, 2003/2005, 2006/2008, and 2009/2011. Using Scopus's own journal classification schema—All Science Classification Codes (ASJC)—journal-to-journal citation data were aggregated into the subject area level. These 27 subject areas were used as the unit of analysis. Journal multi-assignment at the subject area level was considered (i.e., if journal j_1 cited journal j_2 n times and j_1 was assigned to subject s_1 and j_2 was assigned to subject s_2 and s_3, then there are two citation links: one from s_1 to s_2 and one from s_1 to s_3 both with the weight n). The numbers of journals and citations for each ASJC major subject area are shown in Table 1 (C: the number of citations; J: the number of journals).

Table 1. Numbers of journals and incoming citations for ASJC major subject areas

	Subjects		1997/1999	2000/2002	2003/2005	2006/2008	2009/2011
1	General	C	143,556	150,304	157,982	146,292	156,696
		J	59	57	83	98	119
2	Agricultural and Biological Sciences	C	176,256	214,660	281,348	333,983	374,719
		J	1,146	1,207	1,299	1,547	1,564
3	Arts and Humanities	C	7,962	8,712	12,044	13,794	15,422
		J	656	984	1,376	1,565	1,726
4	Biochemistry	C	876,618	1,015,893	1,215,847	1,243,889	1,245,863
		J	1,177	1,319	1,405	1,567	1,558

[2] http://ebrp.elsevier.com/index.asp

5	Business, Management and Accounting	C	14,927	20,195	29,042	40,856	51,430
		J	447	485	674	809	843
6	Chemical Engineering	C	116,588	156,354	248,566	335,412	433,887
		J	409	457	491	486	461
7	Chemistry	C	347,299	469,693	646,849	820,491	979,918
		J	632	674	718	738	729
8	Computer Science	C	40,259	68,180	129,370	158,545	200,468
		J	729	807	973	1,108	1,137
9	Decision Sciences	C	9,049	11,192	17,054	25,813	34,237
		J	136	149	177	212	222
10	Earth and Planetary Sciences	C	81,379	105,332	135,888	155,351	187,988
		J	788	770	860	909	867
11	Economics, Econometrics and Finance	C	15,012	17,261	23,921	32,443	41,509
		J	318	362	471	613	631
12	Energy	C	21,689	29,736	41,576	62,057	104,625
		J	205	241	286	291	291
13	Engineering	C	151,549	218,990	336,940	430,236	561,513
		J	1,558	1,674	1,992	2,008	1,934
14	Environmental Science	C	96,460	123,889	165,832	219,992	300,113
		J	786	799	877	976	999
15	Immunology and Microbiology	C	210,042	240,597	280,186	289,386	289,270
		J	359	371	384	415	404
16	Materials Science	C	211,113	281,410	408,557	521,863	614,036
		J	748	807	926	924	896
17	Mathematics	C	52,334	72,036	116,384	147,284	185,110
		J	639	709	829	978	995
18	Medicine	C	1,199,012	1,508,466	1,883,407	1,966,601	2,052,114
		J	4,093	4,659	5,083	5,516	5,362
19	Neuroscience	C	171,760	204,718	233,529	250,574	252,858
		J	302	319	352	382	379
20	Nursing	C	28,336	41,534	56,870	59,513	69,289
		J	282	310	362	504	487
21	Pharmacology	C	132,571	170,099	215,512	243,774	256,516
		J	433	478	507	556	543
22	Physics and Astronomy	C	317,000	403,280	540,631	615,121	686,994
		J	725	749	853	907	890
23	Psychology	C	50,812	62,253	82,330	101,441	119,384
		J	809	921	1,055	1,209	1,284
24	Social Sciences	C	49,905	61,805	85,652	114,992	141,828
		J	2,081	2,456	3,024	3,562	3,749
25	Veterinary Sciences	C	13,557	17,927	22,181	26,868	32,907

		J	135	144	145	184	179
26	Dentistry	C	7,843	10,990	16,689	21,327	23,945
		J	85	97	108	126	125
27	Health Professions	C	20,299	26,502	34,542	40,072	51,206
		J	281	318	355	405	407
Sum		C	4,563,187	5,712,008	7,418,729	8,417,970	9,463,845
		J*	13,655	14,772	17,559	20,565	23,473
Number of the same journals from the previous period		J	-	11,248	12,700	14,526	16,674
		%	-	76.14%	72.32%	70.63%	71.03%

*Some journals are associated with more than one major subject area; thus the total numbers of unique journals are smaller than the aggregated sum of journals from all 27 major subject areas.

Because different subjects vary greatly in terms of volumes of citations, as can be seen in Table 1, proper normalizations are necessary before the path finding method can be applied. The 27 by 27 subject level citation matrices were processed by normalizing each row independently with a mean of 1 and a standard deviation of 1, a popular approach in network processing (e.g., Flores-Fernández et al., 2012). For each x_{ij} in row *i*, the normalized value x_{ij}^{norm} can be represented by $x_{ij}^{norm} = (x_{ij} - xmean) \times (\frac{ystd}{xstd}) + ymean$ in which $xmean = \frac{x_{ij}}{\sum_j x_{ij}}$, $xstd = \sqrt{\frac{1}{N}\sum_j (x_{ij} - xmean)^2}$, $ymean = 1$, and $ystd = 1$. The effect of different network normalization techniques on path finding is assessed in the *Discussion* section. Figure 1 shows an example of normalizing a 3 by 3 citation network. Each row shows the volume of outgoing citations.

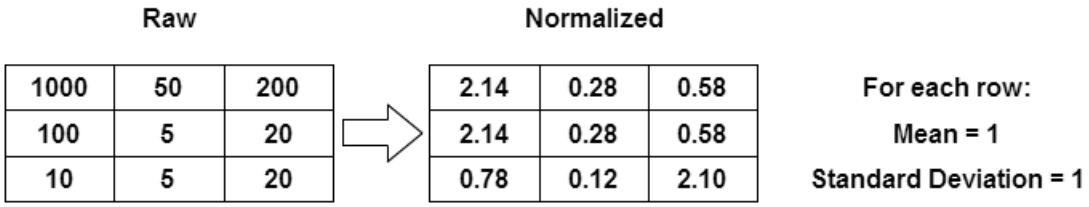

Figure 1. An example of network normalization

Using the trading metaphor (Yan et al., 2013), the normalized subject-level citation links were converted into knowledge flow links such that if subject A has cited subject B with normalized weight *a*, there will be a knowledge flow from the cited subject B to the citing subject A with the same weight of *a* (see an example in Figure 2 (1.a and 1.b)). The normalized weights provide a more accurate investigation of network paths as these values are not susceptible to a high skewness which is present in most citation-based data sets.

Method

Maximum spanning tree

The standard path finding algorithms assume that the citation networks should be: (1) binary, that is, all citation links have the same weight of one; and (2) acyclic, that is, there should be no loop in the network (White, 2003). A typical example of this type of networks is the paper citation networks. However, many

real-world networks are weighted and cyclic: their links have different strengths and they have at least one directed path that starts and ends in the same node. Examples of this type of networks include author citation networks, journal citation networks, and subject citation networks. The standard path finding algorithm, therefore, is not applicable to such networks, because it cannot make use of the information such as link weight.

In light of this, we employ the maximum spanning tree (MST) algorithm to find paths in weighted, directed, and cyclic networks. To define a MST, we first introduce a spanning tree. For network G with vertex set V and link set E, a spanning tree is sub-network that contains all nodes V in G and subset of links from E that connects V such that it becomes a tree (i.e., no cycles). A MST is a spanning tree that is connected by a subset of links with the highest weight to form a tree. MST extracts critical knowledge paths in a network. These paths form the smallest set of links to maintain the flow of information—any further link removal from a MST will result in flow disruptions. For network G with vertex set V and link set E, the procedures of extracting a MST are (Kruskal, 1956):

1. Sort the link set E in G in descending order based on link weight. Form an empty link set S. S will be used to contain links in the MST.
2. Add the first link (i.e., the one with the highest link weight) to S.
3. Add the next link to S if and only if it does not form a cycle in S.
4. If S has $n-1$ links where n is the size of the vertex set E, then return S which is the obtained MST; if all links in E have been traversed and the number of links in S is smaller than n-1, then return with the information that G is a disconnected network; otherwise go back to step 3.

This algorithm works with both directed and undirected networks. In the case of directed networks, a link from node A to B does not guarantee the existence of a reciprocal link from B to A. Assuming the citation flow network in Figure 2 is a normalized network through the method introduced in the *Data* section, to extract a MST in a knowledge flow network, the first step is to convert this normalized citation network by reversing the link directions, because knowledge flows into the citing entity from the cited one. We then use the resulted knowledge flow network to illustrate how to extract a MST from this network:

1. Create an empty link set S. Rank all links in descending order. Add the links with the highest weight to S: A-C and A-D (Figure 2 (2.a)).
2. Then add the link with the next highest weight to S: A-B (Figure 2 (2.b)).
3. Then add the links with the next highest weight to S: C-E, C-F, and D-H (Figure 2 (2.c)).
4. Then add the link with the next highest weight to S: C-G (Figure 2 (2.d)).
5. Then add the links with the next highest weight to S: E-I and F-K. Links E-A and F-G were not added because adding them will result in cycles (Figure 2 (2.e)).
6. Lastly, add the link with the next highest weight to S: E-J. Now S has 10 links which the number of nodes minus one. S is thus returned which contains the MST of the sample network (Figure 2 (2.f)).

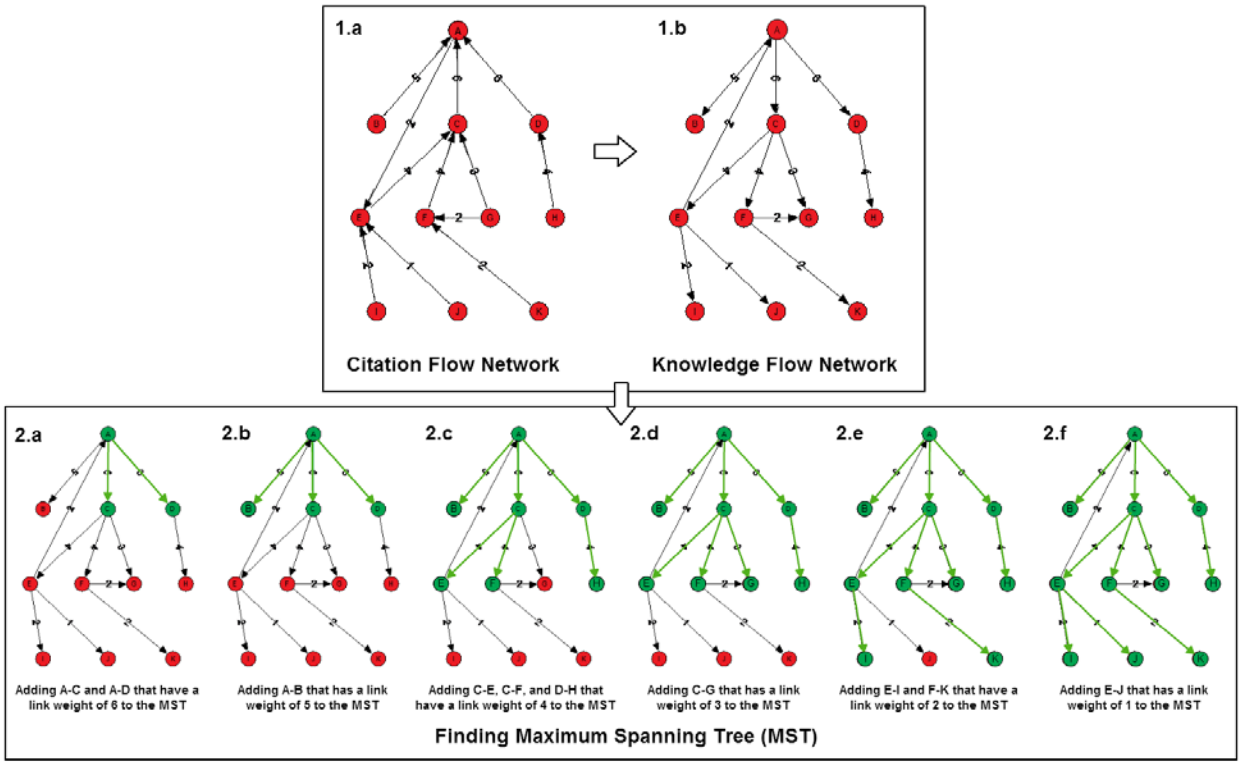

Figure 2. An example of identifying paths through maximum spanning tree

Citation flow, knowledge flow, and knowledge path

In the past, several terms have been used to denote the flow of knowledge, including citation flow, knowledge flow, and knowledge path. They are used interchangeably in some cases whereas distinctively in the others. We give an operational definition of these related terms by using the same sample network in Figure 2.

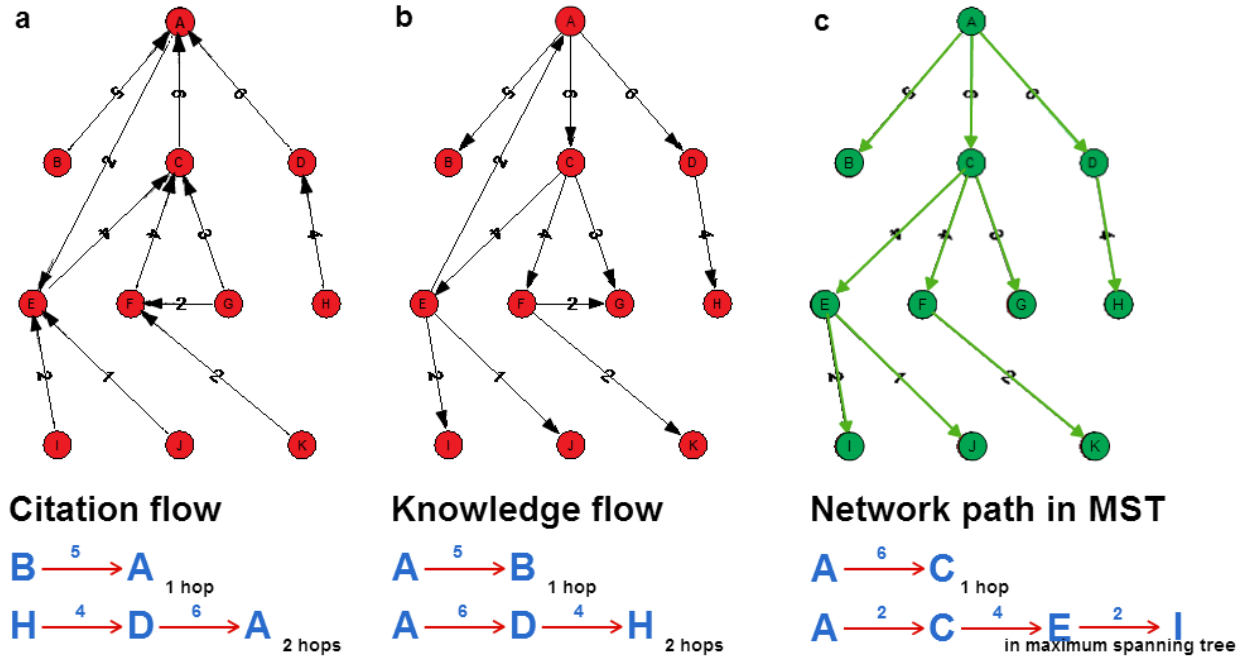

Figure 3. Examples of citation flow, knowledge flow, and knowledge path

Figure 3 (a) is a citation flow network in that a link denotes the flow of citations from the citing node to the cited one, for instance, a one-hop citation flow from B to A and a two-hop citation flow from H to D and then to A. Using the trading metaphor, two knowledge flows are developed: a one-hop knowledge flow from A to B and a two-hop knowledge flow from A to D and then to H (Figure 3 (b)).

Knowledge flows focus on nodes' neighboring link structure; knowledge paths, on the other hand, consider the topology of the network. For instance, shortest path (e.g., Yan, 2014), epidemic models (e.g., Bettencourt et al., 2006; Kiss et al., 2010), maximum spanning tree (e.g., Gomez-Rodriguez, Leskovec, & Krause, 2010), and minimal spanning tree (e.g., Chen & Morris, 2003) are representative approaches to identify knowledge paths. One-hop knowledge paths (e.g., from A to C) are indeed one-hop knowledge flows; nonetheless, not all one-hop knowledge flows are one-hop knowledge paths, because some one-hop knowledge flows are not on the knowledge paths (for instance, the knowledge flow from F to G is not a knowledge path). Knowledge paths do not negate the fact that knowledge flows in both directions. We use the concept of path here to denote the extraction of the most important knowledge channels in knowledge flow networks. We also acknowledge the fact that the most appropriate instrument to study knowledge diffusion should be paper citation relations (e.g., Hethcote, 2000; Kiss et al., 2010). The discipline-level citations may not possess sufficient granularity to warrant a diffusion study because the time stamps for individual papers—the very ingredient for diffusion studies—are aggregated and non-traceable. Thus, realizing this limitation, this study will largely focus on the analysis of the network structure of several disciplinary citation networks through the MST method and will only make tentative implications to knowledge diffusion.

Results

Subjects in the maximum spanning tree

This section presents the results on important subjects in the maximum spanning tree. In the language of graph theory, root subjects are the parent nodes and the branched subjects are the child nodes. The fact that a subject performs a parent role implies that it has an advantage to control and influence its child subjects. Table 2 shows for each subject, the number of subjects that are located in its branches. Thirteen subjects are included in Table 2, because the rest 14 subjects did not possess branches for any citation window.

Table 2. Number of subjects located in each subject's branches

	Number of subjects located in the subject's branches (rank)					
Subjects	1997/1999	2000/2002	2003/2005	2006/2008	2009/2011	All Time
Medicine	26 (1)	26 (1)	26 (1)	26 (1)	26 (1)	130 (1)
Biochemistry	13 (2)	13 (2)	13 (2)	13 (2)	13 (2)	65 (2)
Chemistry	10 (3)	8 (3)	9 (3)	8 (3)	9 (3)	44 (3)
Physics and Astronomy	6 (4)	6 (4)	6 (5)	4 (5)	1 (8)	23 (4)
Materials Science	0 (-)	0 (-)	7 (4)	5 (4)	7 (4)	19 (5)
Social Sciences	3 (5)	3 (5)	3 (7)	3 (7)	3 (6)	15 (6)
Engineering	1 (7)	1 (7)	4 (6)	4 (5)	4 (5)	14 (7)
Agricultural and Biological Sciences	3 (5)	2 (6)	1 (9)	2 (8)	1 (8)	9 (8)
Computer Science	0 (-)	0 (-)	2 (8)	2 (8)	2 (7)	6 (9)
Mathematics	1 (7)	1 (7)	1 (9)	1 (10)	1 (8)	5 (10)
Business, Management and Accounting	0 (-)	1 (7)	1 (9)	1 (10)	1 (8)	4 (11)
Environmental Science	1 (7)	1 (7)	0 (-)	1 (10)	0 (-)	3 (12)
Economics, Econometrics and Finance	1 (7)	0 (-)	0 (-)	0 (-)	0 (-)	1 (13)

The top three subjects with the highest numbers of branches—Medicine, Biochemistry, and Chemistry—maintained stable positions over the past five citation windows. The numbers of branches for Social Sciences and Mathematics also remained steady. In the meantime, some changes can be found: while Materials Science, Engineering, and Computer Science secured more branches, Physics reduced branches from six in the first three windows, to four in 2006/2008, and to one in 2009/2011. Economics used to be the root subject for Business in the first window and became its branch in the recent four windows. The results may be explained by the changes in raw numbers of incoming citation (see Table 1): for instance, for the three related subjects, the growth rates for Engineering and Materials Science are higher than that for Physics; likewise, the number of incoming citations for Business surpassed the number for Economics since the second window. At the same time, we should be aware of the confounding factors that may contribute to this outcome (e.g., reclassification of journals and release of new journals) and be reminded that subdomain analyses are needed so as to avoid misinterpretations of the disciplines' waxing and waning status.

Paths in the maximum spanning tree

This section identifies the one-hop paths with the highest normalized weight in the MST. According to the trading metaphor, volumes of citations denote scientific trading impact. Thus, the paths with the highest normalized weight listed in Table 3 are the ones with the highest trading impact. Twenty-six paths

are needed to connect all subjects for each citation window. There are 35 unique paths when combining the paths from all five MSTs. Among these, 18 paths occurred in all five MSTs; additionally, three occurred in four, four occurred in three, six occurred in two, and four occurred in one MST. The top 20 paths by path weight in Table 3 include all 18 paths that occurred in five MSTs and two paths that occurred in four MSTs (the one from Physics to Engineering and the one from Business to Economics).

Table 3. Top 20 paths with the highest normalized weight in the maximum spanning tree (one-hop)

	Paths with the highest normalized weight in the maximum spanning tree (rank)					
Paths	1997/1999	2000/2002	2003/2005	2006/2008	2009/2011	All Time
Medicine > Health Professions	5.60 (2)	5.62 (2)	5.67 (1)	5.65 (1)	5.59 (1)	28.14 (1)
Medicine > Nursing	5.61 (1)	5.64 (1)	5.64 (2)	5.63 (2)	5.56 (2)	28.08 (2)
Medicine > Pharmacology	4.38 (5)	4.54 (3)	4.67 (3)	4.58 (3)	4.61 (4)	22.78 (3)
Medicine > Neuroscience	4.36 (6)	4.45 (6)	4.54 (4)	4.58 (4)	4.67 (3)	22.61 (4)
Biochemistry > General	4.60 (3)	4.53 (4)	4.43 (6)	4.39 (7)	4.23 (8)	22.18 (5)
Chemistry > Chemical Engineering	4.46 (4)	4.50 (5)	4.41 (7)	4.40 (6)	4.35 (7)	22.12 (6)
Medicine > Immunology	4.28 (8)	4.36 (7)	4.45 (5)	4.44 (5)	4.52 (5)	22.05 (7)
Social Sciences > Arts and Humanities	4.29 (7)	4.19 (8)	4.25 (8)	4.37 (8)	4.43 (6)	21.52 (8)
Medicine > Psychology	3.94 (11)	4.12 (9)	4.11 (10)	4.13 (9)	4.06 (11)	20.37 (9)
Medicine > Dentistry	4.21 (9)	4.10 (10)	4.00 (12)	3.99 (12)	4.07 (10)	20.36 (10)
Medicine > Biochemistry	3.91 (12)	4.01 (11)	4.12 (9)	4.10 (10)	4.20 (9)	20.33 (11)
Medicine > Veterinary Sciences	4.06 (10)	3.91 (12)	4.07 (11)	4.08 (11)	4.03 (12)	20.14 (12)
Mathematics > Decision Sciences	3.89 (13)	3.56 (14)	3.56 (13)	3.74 (13)	3.61 (13)	18.36 (13)
Engineering > Computer Science	3.47 (17)	3.72 (13)	3.01 (20)	3.08 (17)	3.10 (15)	16.38 (14)
Agricultural Sci > Environmental Sci	3.37 (18)	3.25 (18)	3.43 (14)	3.16 (15)	2.79 (21)	16.00 (15)
Biochemistry > Agricultural Sci	3.22 (20)	3.18 (20)	3.10 (18)	2.91 (22)	2.80 (20)	15.21 (16)
Medicine > Social Sciences	2.75 (22)	2.79 (22)	2.94 (21)	2.92 (21)	2.72 (22)	14.12 (17)
Physics > Engineering	3.64 (16)	3.29 (17)	3.23 (16)	3.02 (18)	-	13.17 (18)
Business > Economics	-	2.91 (21)	2.79 (22)	2.92 (20)	2.96 (17)	11.58 (19)
Biochemistry > Chemistry	2.39 (23)	2.27 (24)	2.30 (24)	2.25 (24)	2.15 (25)	11.36 (20)

The path from Medicine to Health Professions has the highest overall normalized weight, indicating a salient flow from the knowledge exporter Medicine to the importer Health Professions. Additionally, there are nine other paths from Medicine to related disciplines among the top 20 paths. These results were obtained from the normalized knowledge flow networks thus suggesting an indispensable role of Medicine in communicating with neighboring disciplines—not only measured by its absolute size but also by the normalized relative importance. Also on the top 20 list include path such as the ones from Chemistry to Chemical Engineering, form Social Sciences to Arts and Humanities, and from Mathematics to Decision Sciences. These paths are among the most important links in the knowledge flow networks: the removal of any will result in the fragmentation of the spanning trees. Overall, the rankings for these top ranked paths stayed stable during the past five citation windows.

Maximum spanning tree and its dynamics

This section reports the results of the MSTs extracted from the knowledge flow networks. Because most real-world networks are complex networks (Barabási, 2002), appropriate clustering and mapping techniques are required before these networks are sensible to the audience. Clustering and mapping sometimes can be conducted independently, but can also be achieved in an integrated way, such as the VOSviewer Clustering and Mapping technique (Waltman, van Eck, & Noyons, 2010). The subject-level knowledge flow networks, measured by the number of nodes, are by no means a complex network; thus clustering may not be a concern here. However, these networks are quite dense, because almost every node is connected with one another through citations. Consequently, it is not feasible to illustrate all knowledge flows without using certain complexity deduction method. To effectively capture essential knowledge flows, we utilize the built-in link reduction component in MST and only show the paths on the spanning tree. The spanning tree efficiently reduces the complexity of dense networks and codifies the "backbone of science" (Boyack, Klavans, & Börner, 2005, p. 351). The normalized weight for each of the 26 paths in the MST is also shown in Figure 4.

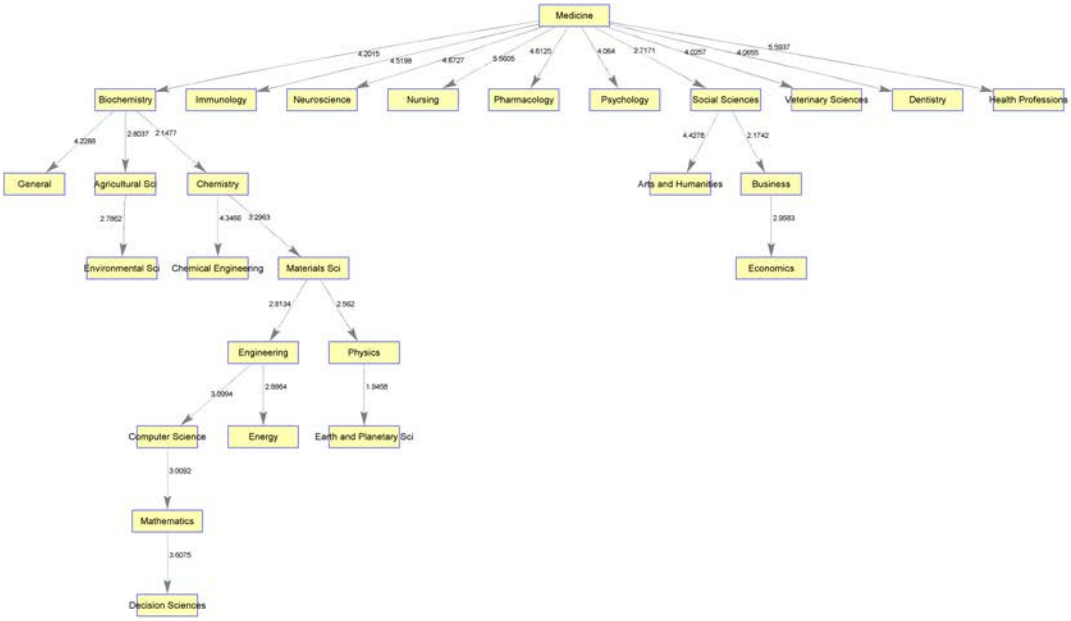

Figure 4. Maximum spanning tree extracted from the 2009/2011 knowledge flow network

Medicine is the root of the MST and has 10 first-level branches because it served as the largest exporter of knowledge for these 10 subjects (as measured by the normalized numbers of citations). Among them, Biochemistry and Social Sciences have second-level branches. The Biochemistry-branch is the most elaborate: it has three second-, three third-, two fourth-, three fifth-, one sixth-, and one seventh-level branches. The Social Sciences branch has two second- and one third-level branches. Several multi-hop knowledge paths can be identified from Figure 4:

- Medicine>>>Biochemistry>>>Agriculture Sciences>>>Environmental Science;
- Medicine>>>Biochemistry>>>Chemistry>>>Materials Science>>>Physics>>>Earth and Planetary Science;
- Materials Science>>>Engineering>>>Computer Science>>>Mathematics>>>Decision Sciences;
- Social Sciences>>>Business>>>Economics.

Figure 5 illustrates the topology changes of the MST across the five citation windows. Locations of the first-level branch remained stable during the five citation windows.

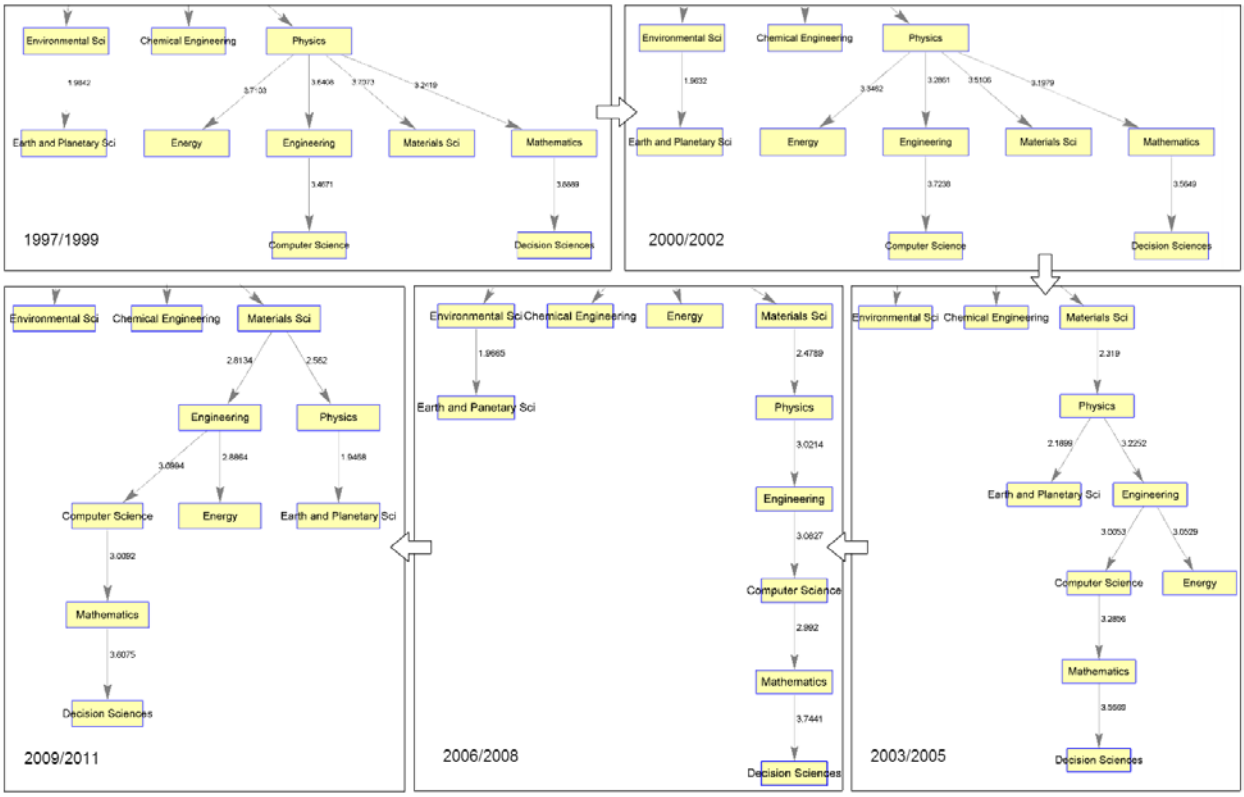

Figure 5. Topology changes of the MST

Figure 5 shows that the locations of the following subjects have changed in the MST:

- Earth and Planetary Sciences: this subject was a branch of Environmental Science with the exception that it was a branch of Physics in 2003/2005 and 2006/2008.
- Physics and Engineering: Engineering was a branch of Physics from 1997/1999 to 2006/2008 and was then in the same level as Physics in 2009/2011.
- Physics and Materials Science: Materials Science was a branch of Physics in 1997/1999 and 2000/2002 and was then the parent node of Physics in 2006/2008 and 2009/2011.
- Mathematics and Computer Science: Computer Science was a branch of Mathematics in 1997/1999; they were located in two branches in 2000/2002 (Mathematics was in the fifth-level and Computer Science was in the sixth-level); Computer Science became the parent node of Mathematics in the recent three windows.
- Business and Economics (not shown in Figure 5): Business was a branch of Economics in 1997/1999 and a parent node for Economics since 2000/2002.

These topological changes may indicate a shift balance within these subjects in trading knowledge: the fact that some subjects (e.g., Materials Science, Computer Science, and Business) may have offered knowledge that others considered as useful and gradually procured more branches in MSTs; however, there may be alternative explanations and the data set of this study may not suffice a conclusive diagnosis

of these changes. Fine-grained subfield analyses are necessary to further interpret these phenomena which will be pursued as a future research direction.

The robustness of the MST method for networks normalized using different techniques is lastly investigated. Different choices of *ymean* and *ystd* in $x_{ij}^{norm} = (x_{ij} - xmean) \times (\frac{ystd}{xstd}) + ymean$ will not affect the network topology because the formula shows that the normalized weight will change in proportion to different *ymean* and *ystd* values. Another popular network normalization method is to divide the cell value by the sum value of the corresponding row $x_{ij}^{norm\prime} = \frac{x_{ij}}{\sum_j x_{ij}}$ such that the sum value of each row is one. The MST method was also applied to networks normalized using this technique and results showed that the topology of the spanning trees was identical to that in Figures 4 and 5 thus providing evidence to support the robustness of the MST method. It is possible that other network normalization techniques may affect the resulted tree topology; nonetheless, we speculate that the variances will only occur at local branches.

Discussion

Science maps

This section discusses the obtained MST with related science maps. To date, more than 20 types of science maps have been proposed (e.g., Boyack, Klavans, & Börner, 2005; Bollen et al., 2009; Klavans & Boyack, 2009; Leydesdorff & Rafols, 2009; Rafols, Porter, & Leydesdorff, 2010; Van Eck & Waltman, 2010; Börner et al., 2012). These maps are mostly co-occurrence-based, using journal co-citation, journal bibliographic coupling, or journal clickstream data. The goal of these maps is to measure the cognitive proximity of one field to another. Klavans and Boyack (2009) have merged 20 existing science maps and identified the general proximity patterns of disciplines (i.e., the consensus map): starting from mathematics, there are "physics, physical chemistry, engineering, chemistry, earth sciences, biology, biochemistry, infectious diseases, medicine, health services, brain research, psychology, humanities, social sciences, and computer science" (p. 455). Compared with this consensus map, the obtained MST exhibits similar characteristics, for instance, in Figure 4, Medicine and Biochemistry are connected, so are Computer Science and Mathematics. Nonetheless, at the network-level, discrepancies can be found: in the consensus map, the general sequence is from mathematics to physics, to chemistry, to earth science, to medicine, and then to social science; however, one of the paths in the MST is from medicine to materials science to engineering, to computer Science, and to mathematics. Moreover, the most defining difference is that the consensus map is a co-occurrence-based map; thus orders are not decisive. For the MST method, the spanning tree denotes the direction of knowledge flows.

A recent development of the flow map (Rosvall & Bergstrom, 2008) is probably the most similar to what we intend to attain here. In the flow map, molecular & cell biology and medicine are in the center connecting physical and life sciences; two regional hubs are also present: physics powers physical sciences (e.g., mathematics, engineering, and geosciences) and economics powers social sciences (e.g., political science, education, sociology, and business). From a visualization perspective, this flow map resembles the above mentioned co-occurrence-based science maps with the goal to partition journals or fields into groups. An examination of knowledge paths is still latent from these maps.

Through MST, we have identified network paths from small but highly dense networks. MST illustrates the structure of the disciplinary knowledge flow networks: subjects such as Medicine, Biochemistry, and Social Sciences are in the front of the paths and tend to acquire a higher number of branched subjects. If we further classify these 27 subjects using the notion of STEM fields, we observed that the two critical links between STEM and non-STEM fields are from Mathematics to Decision Science and from Medicine to Social Sciences. The emerging trend of science teams has the potential to promote cross-field knowledge transfer as teams may comprise scientists and scholars of various expertise and backgrounds (e.g., Moody, 2004; Wuchty, Jones, & Uzzi, 2007).

Applicability issues of a local path finding method

In addition to the MST method employed in this paper, we also attempted to use a local path finding (LPF) method. LPF started off with a selected subject and identified the next connected subject that imports the highest amount of knowledge until there are no untraversed subjects left. However, we found that LPF is not applicable to diffusion studies: because LPF can only take one dependent at a time, the path will be largely rearranged if the weight of two closely valued links has changed in different citation windows. For instance, for links A-B and A-C, if the weight for A-B was slightly higher than A-C in time t but slightly smaller than A-C in time t+1, then A-B will be chosen by LPF in time t and A-C will be chosen by LPF in time t+1. These minor link weight changes interrupted the whole linear chain because of the one dependent rule. Therefore, we believe that LPF is susceptible to subtle changes in scores and is not a robust method for this type of knowledge diffusion study.

Limitations

The limitations of this study are largely derived from the characteristics of the citation data set. First, the journal-level citations were aggregated into the discipline level based on an artificial journal classification scheme ASJC. Although this scheme has an extensive index of all fields of sciences and social science, its coverage may be biased toward certain fields (see Table 1) and it is simply used as a proxy to study disciplinary knowledge flows. Second, the unit of analysis of this paper is disciplines and this unit has limited us from drawing firm conclusions on patterns of knowledge diffusion. Granular paper unit analyses should be able to shed light on research in this direction. Third, the data set only possessed plain citation and publication information and it lacked rich features to help gain more insights from the citation data. To reveal mechanisms of the changes in the network structure, additional data sources are expected to supplement the analyses; for instance, funding policies, community evolving features, and other socio-technical factors can be cross-referenced to understand reasons that may lead to the dynamic changes.

Conclusion

In this paper, we have employed the MST algorithm to a dynamic discipline-level citation data set with the goal to identify knowledge paths among scientific disciplines. The use of the MST algorithm has helped gain understanding of the characteristics of disciplinary citation networks at three levels, including subjects, knowledge paths, and knowledge flow networks.

This study has found that subject such as Medicine, Biochemistry, Chemistry, Materials Science, Physics, and Social Sciences are the ones with multiple branches in the spanning tree and thus may be associated with the ability to influence their branches. In addition, this study has found that most paths connect

STEM fields (e.g., Medicine>>>Biochemistry>>>Agriculture Sciences>>>Environmental Science and Medicine>>>Biochemistry>>>Chemistry>>>Materials Science>>>Physics>>>Earth and Planetary Science); two critical links between STEM and non-STEM fields are the one from Mathematics to Decision Science and the one from Medicine to Social Sciences.

Diachronically, this study has found subjects such as Business, Materials Science, Engineering, and Computer Science have moved toward the root of the spanning tree. For network paths, this study has helped identify the shifting impact between Physics and Engineering, between Physics and Materials Science, between Mathematics and Computer Science, and between Business and Economics. Future research will benefit from conducting fine-grained subfield analyses using rich multi-facet to reveal the latent mechanisms that lead to the dynamic changes.

Acknowledgement

The data set used in this paper is supported by the Elsevier Bibliometric Research Program (EBRP). Qi Yu was supported by the National Natural Science Foundation of China as part of the project "Cooperation Analysis of Technology Innovation Team Member Based on Knowledge Network – Empirical Evidence in the Biology and Biomedicine Field" (No. 71103114) and the project "The Theoretical and Empirical Study of S&D Collaboration Characterized by Collaboration Prediction – Based on the Data in Biomedical Fields" (No. 71473154).

References

Almeida, P., & Kogut, B. (1999). Localization of knowledge and the mobility of engineers in regional networks. *Management Science, 45*(7), 905-917.

Autant-Bernard, C., Mairesse, J., & Massard, N. (2007). Spatial knowledge diffusion through collaborative networks. *Papers in Regional Science, 86*(3), 341-350.

Bacchiocchi, E., & Montobbio, F. (2009). Knowledge diffusion from university and public research. A comparison between US, Japan and Europe using patent citations. *Journal of Technology Transfer, 34*(2), 169-181.

Barabási, A. L. *Linked: How everything is connected to everything else and what it means for business, science, and everyday life*. New York: Penguin Group.

Bettencourt, L. M. A., Kaiser, D. I., Kaur, J., Castillo-Chávez, C., & Wojick, D. E. (2008). Population modeling of the emergence and development of scientific fields. *Scientometrics, 75*(3), 495-518.

Black, D. (1976). Partial justification of the Borda count. *Public Choice, 28*(1), 1-15.

Bollen, J., Van de Sompel, H., Hagberg, A., Bettencourt, L., Chute, R., Rodriguez, M. A., & Balakireva, L. (2009). Clickstream data yields high-resolution maps of science. *PLoS One, 4*(3), e4803.

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., ... & Boyack, K. W. (2012). Design and update of a classification system: the UCSD map of science. *PloS One, 7*(7), e39464.

Börner, K., Penumarthy, S., Meiss, M., & Ke, W. (2006). Mapping the diffusion of scholarly knowledge among major U.S. research institutions. *Scientometrics, 68*(3), 415-426.

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

Buter, R. K., Noyons, E. C. M., & Van Raan, A. F. J. (2010). Identification of converging research areas using publication and citation data. *Research Evaluation, 19*(1), 19-27.

Chen, C., & Hicks, D. (2004). Tracing knowledge diffusion. *Scientometrics, 59*(2), 199-211.

Chen, C., & Morris, S. (2003). Visualizing evolving networks: Minimum spanning trees versus pathfinder networks. In *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on* (pp. 67-74). IEEE.

Chen, C., Leydesdorff, L. (2014). Patterns of connections and movements in dual-map overlays: A new method of publication portfolio analysis. *Journal of the American Society for Information Science and Technology, 65*(2), 334-351.

Cohen, W. M., Nelson, R. R., & Walsh, J. P. (2002). Links and impacts: The Influence of public research on industrial R&D. (In A. Geuna, A. J. Salter, & W. E. Steinmueller, Eds.) *Management Science, 48*(1), 1-23.

Cronin, B. & Pearson, S. (1990). The export of ideas from information science. *Journal of Information Science, 16*(6), 381-2391.

Cronin, B., & Davenport, L. (1989). Profiling the professors. *Journal of Information Science, 15*(1), 13-20.

David, P. A., & Foray, D. (2002). An introduction to the economy of the knowledge society. *International Social Science Journal, 54*(171), 9-23.

Dummett, M. (1998). The Borda count and agenda manipulation. *Social Choice and Welfare, 15*(2), 289-296.

Flores-Fernández, J. M., Herrera-López, E. J., Sánchez-Llamas, F., Rojas-Calvillo, A., Cabrera-Galeana, P. A., Leal-Pacheco, G., ... & Martínez-Velázquez, M. (2012). Development of an optimized multi-biomarker panel for the detection of lung cancer based on principal component analysis and artificial neural network modeling. *Expert Systems with Applications, 39*(12), 10851-10856.

Geroski, P. A. (2000). Models of technology diffusion. *Research policy, 29*(4), 603-625.

Gomez-Rodriguez, M., Leskovec, J., & Krause, A. (2010). Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1019-1028). New York: ACM Press.

Hethcote, H. W. (2000). The Mathematics of Infectious Diseases. *SIAM Review, 42*(4), 599-653.

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.

Jaffe, A., & Trajtenberg, M. (1999). International knowledge flows: Evidence from patent citations. *Economics of Innovation and New Technology, 8*, 105-136.

Kiss, I. Z., Broom, M., Craze, P. G., & Rafols, I. (2010). Can epidemic models describe the diffusion of topics across disciplines? *Journal of Informetrics, 4*(1), 74-82.

Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for information science and technology, 60*(3), 455-476.

Knorr-Cetina, K. (1999). *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society, 7*(1), 48-50.

Larivière, V., Sugimoto, C. R., & Cronin, B. (2012). A bibliometric chronicling of Library and Information Science's first hundred years. *Journal of the American Society for Information Science and Technology, 63*(5), 997-1016.

Lewison, G., Rippon, I., & Wooding, S. (2005). Tracking knowledge diffusion through citations. *Research Evaluation, 14*(1), 5-14.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology, 60*(2), 348-362.

Lockett, A., & McWilliams, A. (2005). The balance of trade between disciplines: do we effectively manage knowledge? *Journal of Management Inquiry, 14*(2), 139-150.

MacGarvie, M. (2005). The determinants of international knowledge diffusion as measured by patent citations. *Economics Letters, 87*(1), 121-126.

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.

Ponds, R., Van Oort, F., & Frenken, K. (2007). The geographical and institutional proximity of research collaboration. *Papers in Regional Science, 86*(3), 423-443.

Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology, 61*(9), 1871-1887.

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences, 105*(4), 1118-1123.

Singh, J. (2005). Collaborative networks as determinants of knowledge diffusion patterns. *Management science, 51*(5), 756-770.

Stigler, S. M. (1994). Citation patterns in the journals of statistics and probability. *Statistical Science, 9*(1), 94-108.

Szulanski, G. (1996). Exploring internal stickiness: Impediments to the transfer of best practice within the firm. *Strategic Management Journal, 17*(1), 27-43.

U.S. Census Bureau (2012). Science & technology: Expenditures, research development. Retrieved September 9, 2013 from http://www.census.gov/compendia/statab/cats/science_technology/expenditures_research_development.html

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523-538.

Van Leeuwen, T., & Tijssen, R. (2000). Interdisciplinary dynamics of modern science: analysis of cross-disciplinary citation flows. *Research Evaluation, 9*(3), 183-187.

Vitanov, N. K., & Ausloos, M. R. (2012). Knowledge epidemics and population dynamics models for describing idea diffusion. In *Models of Science Dynamics*(pp. 69-125). Springer Berlin Heidelberg.

Wagner, C. S., & Leydesdorff, L. (2009). Network structure, self-organization and the growth of international collaboration in science. *Research Policy, 34*(10), 1608-1618.

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

White, H. D. (2003). Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and Technology, 54*(5), 423-434.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5827), 1036-1039.

Yan, E. (2014). Finding knowledge paths among scientific disciplines. *Journal of the American Society for Information Science & Technology, 65*(11), 2331-2347.

Yan, E., Ding, Y., Cronin, B., & Leydesdorff, L. (2013). A bird's-eye view of scientific trading: Dependency relations among fields of science. *Journal of Informetrics, 7*(2), 249-264.