Research Dynamics, Impact, and Dissemination: A Topic-Level Analysis

Erjia Yan
College of Computing and Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA. E-mail: erjia.yan@drexel.edu

In informetrics, journals have been used as a standard unit to analyze research impact, productivity, and scholarship. The increasing practice of interdisciplinary research challenges the effectiveness of journal-based assessments. The aim of this article is to highlight topics as a valuable unit of analysis. A set of topic-based approaches is applied to a data set on library and information science publications. Results show that topic-based approaches are capable of revealing the research dynamics, impact, and dissemination of the selected data set. The article also identifies a nonsignificant relationship between topic popularity and impact and argues for the need to use both variables in describing topic characteristics. Additionally, a flow map illustrates critical topic-level knowledge dissemination channels.

Introduction

Informetric studies primarily focus on two threads of research. One employs journals as the unit of analysis: A paper is published in a journal, and a journal is assigned to certain subject areas or research domains. The other thread is based on authorship: A paper is written by an author, an author is affiliated to an institution, and an institution is located in a certain geographical area. For the first thread, journals have been used as the *de facto* unit to address questions related to impact assessment (e.g., Pinski & Narin, 1976; Bollen, Rodriguez, & Van de Sompel, 2006), clustering and mapping (e.g., Ding, Chowdhury, & Foo, 2000; Van Eck & Waltman, 2010), and scientific trading (e.g., Borgman & Rice, 1992; Leydesdorff & Probst, 2009).

Whereas journals serve as a fixed research instrument to study scholarship and design policies (e.g., Holton, 1978; Van Raan, 2004), science is becoming more collaborative and interdisciplinary (Berners-Lee, Hall, Hendler, Shadbolt, & Weitzner, 2006; Metzger & Zare, 1999). Likewise, the effectiveness of journal-based analyses has been challenged (e.g., Rafols & Leydesdorff, 2009; Waltman & Van Eck, 2012). The prevalence of multidisciplinary, open access journals and online social media (e.g., Mendeley and CiteULike) has further intensified this issue (Björk et al., 2010; Evans & Reimer, 2009).

To reconcile this tension, scholars are calling for a more fine-grained analysis to classify research and assess impact. A set of solutions has been proposed: It includes the use of clustering techniques (e.g., Waltman & Van Eck, 2012; Boyack & Klavans, 2014) and probabilistic topic models (e.g., Blei, Ng, & Jordan, 2003; Blei & Lafferty, 2007; Ramage, Hall, Nallapati, & Manning, 2009) to identify research specialties or topics from papers. These studies focused largely on delineating topics and evaluating algorithmic performances. Consequently, we have a limited understanding of topic-level impact distribution and knowledge dissemination. A topic-level analysis enables research on science dynamics, influence assessment, and knowledge diffusion, filling the gap left by previous research.

The central research question is how to use topic-based approaches to explore research dynamics, impact, and knowledge dissemination. Using a data set on library and information science (LIS) publications, the following questions will be addressed:

- What are the dynamic characteristics of topic popularity and impact in LIS?
- What topics are more self-contained? Do popular topics tend to be highly cited?
- What are the knowledge dissemination patterns of topics in LIS?

This study aims to expand the landscape of informetric research by exemplifying topics as a valuable unit of analysis. Although the findings are only applicable to LIS, the approaches and methods proposed in this paper are domain independent and should inform the studies of topic-level popularity and impact across different fields.

Received January 12, 2014; revised April 22, 2014; accepted April 22, 2014

 DOI: 10.1002/asi.23324

Literature Review

Identifying Research Specialties Through Network Methods

Bibliometric networks have been used to identify research specialties long before the recent proliferation of network analyses. Based on the level of analysis, bibliometric networks can be examined from paper, author, and journal levels. The clustering of papers is made possible through paper cocitation networks (Small, 1973) and paper bibliographic coupling networks (Kessler, 1963). The clustering results have been used to compare the topic similarity of papers (Kessler, 1963; Small, 1973) and to find intellectual turning points (Chen, 2004, 2006). Because cocitation relations reveal research specialties, scholars have used multidimensional scaling (e.g., White & McCain, 1998), factor analysis (e.g., White & Griffith, 1981), and pathfinder networks (e.g., White, 2003) to identify the subdivisions of research domains represented by author cocitation relations. At the journal level, subject categories provide a fixed and consistent journal classification system, but the accuracy of this classification system has been questioned in recent years (e.g., Boyack, Klavans, & Börner, 2005; Rafols & Leydesdorff, 2009). Accordingly, alternative journal classification schemes have been proposed through the use of journal-level bibliometric networks (e.g., Glänzel & Schubert, 2003; Leydesdorff & Vaughan, 2006; Zhang, Liu, Janssens, Liang, & Glänzel, 2010).

In addition to these co-occurrence-based approaches, there is a trend in bibliometrics to use hybrid networks to identify research specialties (e.g., Liu et al., 2010; Janssens, Glänzel, & De Moor, 2008; Boyack & Klavans, 2010; Zitt, Lelu, & Bassecoulard, 2011). It has been shown that hybrid approaches yielded more accurate clustering results than the use of only one type of network (e.g., Boyack & Klavans, 2010; Janssens et al., 2008). The limitation of these network-based analyses is that they rely on the interpretation of author or journal clusters. Thus, they may not provide sufficiently detailed information for a fine-grained analysis of the cognitive structure of research fields. This is largely attributed to the fact that authors or journals may represent a mixture of different specialties.

Identifying Research Specialties Through Topic Models

Topic models have garnered much attention in recent years. We focus on two aspects: citation-aware topic models (models that consider cited references of text corpora) and dynamics topic models (models that are capable of distilling dynamic features from text corpora). Latent Dirichlet allocation (LDA) models each document as a mixture of probabilistic topics. Through LDA, each topic is defined as a multinomial distribution over words (e.g., Blei et al., 2003; Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004; Steyvers, Smyth, Rozen-Zvi & Griffiths, 2004; Blei & Lafferty, 2007; Tang et al., 2008; Tu, Johri, Roth, & Hockenmaier, 2010). These topic models treat a document as a bag-of-words without considering its cited references and can be referred to as citation-unaware models. Citation-aware models, on the other hand, used probabilistic latent semantic analysis or LDA as the generative process on both texts and citations (e.g., Cohn & Hofmann, 2000; Erosheva, Fienberg, & Lafferty, 2004; Dietz, Bickel, & Scheffer, 2007; Nallapati, Ahmed, Xing, & Cohen, 2008; He et al., 2009; Yang, Yin, & Davison, 2011; Masada & Takasu, 2012).

The dynamic aspect of topics has been investigated primarily through three approaches: post-hoc analysis (e.g., Griffiths & Steyvers, 2004; Hall, Jurafsky, & Manning, 2008); segmented approaches (e.g., Bolelli, Ertekin, Zhou, & Giles, 2009); and continuous-time models (Wang & McCallum, 2006). Whereas the post hoc method used topic-document probability distributions to determine topic dynamics, segmented approaches divided document corpora into segments that have contingent time stamps. Both methods rely on the Markov assumption that the state at a single time point is independent from others. Wang and McCallum (2006) found that their non-Markov continuous-time model provided better prediction and more interpretable topical trends.

Data and Methods

Data Set

The field of LIS was chosen as a sample domain to address the proposed research questions. Publications in the *Journal Citation Report* subject category of Information Science and Library Science were harvested as the data set. The following filters were used: (a) year of publication: between 1955 and 2013; (b) document type: article, review article, or proceeding paper; (c) language: English; and (d) records that have no author, title, or cited reference were removed. The intermediate data set comprised 51,156 publications.

Because the LDA model needs *meaningful* words to conduct the inference, further data preprocessing of publications' titles was implemented. To achieve this, (a) title words that have less than three letters (i.e., single- and double-letter words) were removed from titles; (b) 10 most frequently occurred words were removed from titles; (c) those words that occurred in less than three publications were also removed; and, finally, (d) publications whose titles have less than three words were removed from the data set. The final resulted data set comprised 47,137 publications.

The LDA Model

The LDA model was proposed by Blei et al. (2003) to identify topics from large publication corpora. The model assumes that words for each paper are derived from a mixture of topics and each topic follows a multinomial distribution and is defined on a collection of vocabulary. The basic procedures are (Wang & Blei, 2011, p. 450):

DOI: 10.1002/asi

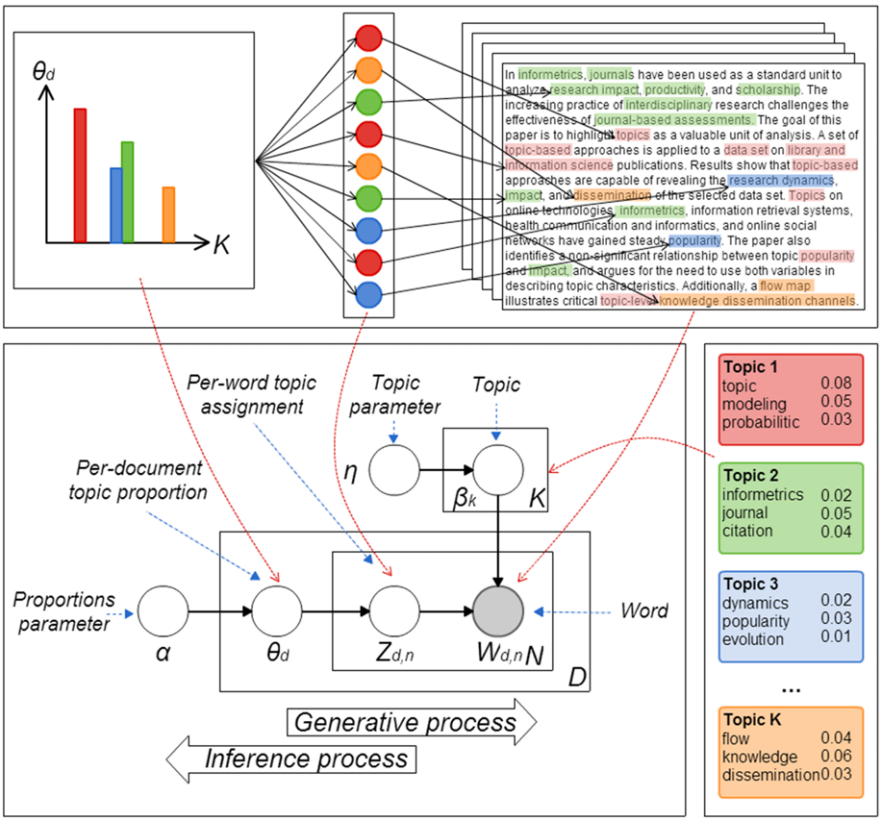

FIG. 1. A graphic model presentation of LDA (Blei, 2012). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

For each document in the data set:

- Draw topic proportions θ_d~Dirichlet (α)
- For each word n
- Draw topic assignment $z_{d,n}$~Mult(θ_d)
- Draw word $w_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})$

θ_d is the per-document topic proportion for the dth document; $z_{d,n}$ is the per-word topic assignment for the nth word in the dth document; and $w_{d,n}$ is the nth word in the dth document. To better understand LDA, we refer to the graphic model with plate notations (Blei, 2012, pp. 78, 81) using the Abstract of the current paper as an example: N in Figure 1 denotes the number of words in a document, D denotes the number of documents, and K denotes the number of topics.

The graphic model shows the *generative process* of LDA. It assumes that each document comprises multiple topics in different proportions (θ_d~Dirichlet (α)); for instance, the current paper is mainly about topic modeling (high proportion), but also covers research dynamics (medium proportion), citation analysis (medium proportion), and knowledge diffusion (medium proportion). These topics are then derived from the per-document topic proportions ($z_{d,n}$~Mult(θ_d)), and words in the document are drawn from these topics ($w_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})$). The generative process is an "imaginary random process by which the model assumes the documents arose" (Blei, 2012, p. 78). Words in a document are the only observed variable (highlighted in Figure 1); per-word topic assignment, per-document topic proportions, and topics are the latent variables to be identified. To achieve this, we rely on the *inference process*. This process uses the observed variable in the joint distribution to compute the conditional distributions (Blei, 2012, p. 80):

$$p(\beta,\theta,z,w)=\left(\prod_{i=1}^{K}p(\beta_i\mid\eta)\right)\left(\prod_{d=1}^{D}p(\theta_d\mid\alpha)\right)\left(\prod_{n=1}^{N}p(z_{d,n}\mid\theta_d)\,p\left(w_{d,n}\mid\beta_{1:K,z_{d,n}}\right)\right).$$

For extensive discussions on computing the posterior distribution from the joint distribution, we refer to Steyvers and Griffiths (2007), Blei and Lafferty (2009), and Blei (2012).

One advantage of the LDA model is that the words of each document are predicated on a mixture of topics, allowing a document to be assigned into multiple topics. This multiassignment strategy delivers more flexible estimations than the classic cocitation analysis (e.g., White & McCain, 1998) or distance-based clustering (e.g., Yan, Ding, & Jacob, 2012). Limitations of the LDA model, however, arise from its rigid assumptions (Blei, 2012): The model assumes that the order of the words in the document is not relevant, nor is the order of documents in the data set, and the model also assumes that the number of topics is known and fixed.

DOI: 10.1002/asi

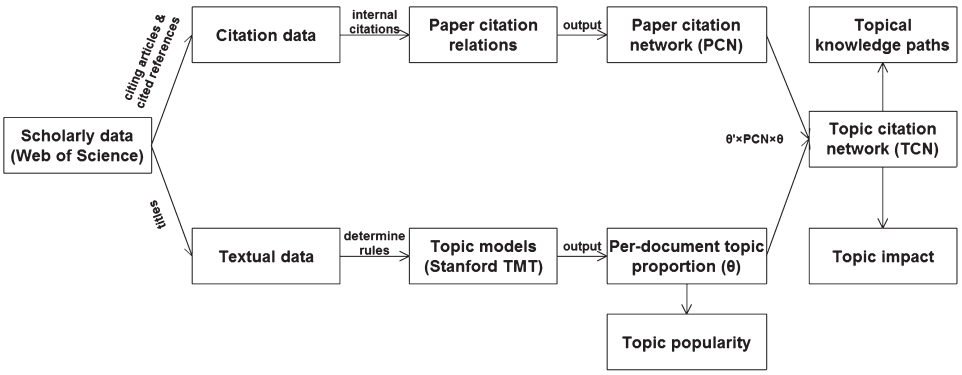

FIG. 2. A flow chart showing the steps to obtain topic popularity and impact.

These assumptions help improve computational efficiency, but they also cause complications when a text corpus does not comply with these assumptions. Solutions to relax these assumptions have been provided (see Blei, 2012). Previous studies on the cognitive structure of LIS have employed a variety of approaches and identified different numbers of research clusters, ranging from a few to more than a dozen (e.g., Åström, 2007; Järvelin & Vakkari, 1993; Prebor, 2010; Sugimoto, Li, Russell, Finlay, & Ding, 2011; Yan, 2014a; Zhao & Strotmann, 2014). Considering these scholarly efforts in relation to the size, diversity, and duration of the current data set, the number of topics was set at 50 for this study. This allows for a detailed diachronical analysis of the cognitive landscape of LIS. The result is evaluated in two ways: first, through Jensen-Shannon divergence (JSD); second, through a qualitative comparison of clusters obtained from previous studies (see Discussion section).

Topic Popularity and Impact

We use Figure 2 to show the major steps to calculate topic popularity and impact. We started by downloading scholarly data from academic databases. The data set was then prepared in two ways. One involved the use of citing articles and cited references to construct a paper citation network (PCN); the other took publication titles and used these textual data as the input for topic modeling. The raw input data were then processed based on certain predefined rules as to what words should be processed by a certain topic model. The Stanford Topic Modeling Toolbox (Stanford TMT: http://nlp.stanford.edu/downloads/tmt/tmt-0.4/) was used to perform the topic modeling. Other recognized toolkits include MALLET (http://mallet.cs.umass.edu/topics.php) and a series of open-source software packages by the Blei group at Princeton (https://www.cs.princeton.edu/~blei/topicmodeling.html). One output file from these topic model implementations is the document-topic proportion file (or simply the θ file). By aggregating θ for each topic, we are informed by the topic popularity; by multiplying θ with PCN ($\theta' \times PCN \times \theta$), we can obtain a topic citation network—this, in turn, informs us of the topic impact and topical knowledge paths. The way to calculate topic popularity and impact as well as topical knowledge paths is elaborated in the following paragraphs.

	t1	t2	t3	t4	t5
p1	0.60	0.03	0.12	0.20	0.05
p2	0.08	0.22	0.13	0.49	0.08
p3	0.33	0.09	0.29	0.17	0.12
popularity	*1.01*	*0.34*	*0.54*	*0.88*	*0.25*

FIG. 3. An example of topic proportions for three papers. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Topic popularity was calculated through θ_d, the per-document topic proportion for document d. It shows that if the topic proportion for a certain topic is high among a number of papers, this topic is considered as popular (e.g., Griffiths & Steyvers, 2004). Figure 3 shows an example of assigning three papers into five topics. As illustrated, the popularity of each topic $Pop(t(j))$ can be calculated through aggregating $\theta_{d,t(j)}$ for each topic t(j).

The topic popularity score $Pop(t(j))$ for topic j was calculated through aggregating $\theta_{d,t(j)}$:

$$Pop(t(j)) = \sum_{d} \theta_{d,t(j)}$$

where $t(j)$ denotes jth topic and $\theta_{d,t(j)}$ denotes the per-document proportion of document d for topic $t(j)$. The topic popularity for topic j in year t can be expressed as:

$$Pop(t(j), t) = \sum_{d|py(d)=t} \theta_{d,t(j)}$$

where $py(d)$ denotes the publication year of document d.

For an effective comparison of topic popularity across different years, $Pop(t(j), t)$ was normalized by the sum of popularity scores of all topics for year t (i.e., the total number of publications of year t):

DOI: 10.1002/asi

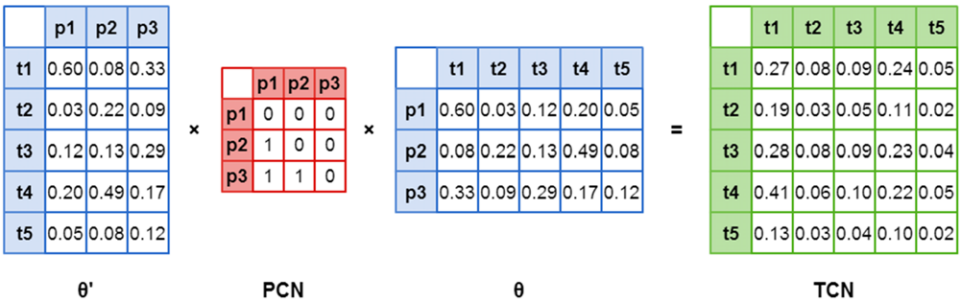

FIG. 4. The formation of a topic citation network with five topics. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$$N\_Pop(t(j),t)=\frac{\sum_{d|py(d)=t}\theta_{d,t(j)}}{\sum_{j}\sum_{d|py(d)=t}\theta_{d,t(j)}}$$

Using $N\_Pop(t(j), t)$ as the dependent variable and year of publication as the independent variable, a single-factor linear regression model can be formed. Slopes of the modeled regression curve were used to measure topic popularity trends.

For each topic, the average year of publication informs us of its development stages. It was calculated by multiplying $\theta_{d,t(j)}$ with $py(d)$, divided by the aggregated sum of $\theta_{d,t(j)}$. The average year of publication of topic j can thus be calculated:

$$py(t(j))=\frac{\sum_{d}(\theta_{d,t(j)}\times py(d))}{\sum_{d}\theta_{d,t(j)}}$$

Slopes of topic popularity can be further evaluated using the z-score (Yan, 2014a), from which labels of declining, fluctuating, and raising topics are assigned. The focus of this paper goes beyond the assignment of topics, to an examination of the relationship between topic popularity and impact. In the following paragraphs, we introduce a way to add citations to topics and form a topic citation network.

An internal PCN was constructed. This citation network comprised 47,137 citing papers and their associated citation relations. The total number of internal citations was 93,932. This is the number of times papers in the network have been cited by other papers in the network. PCN will afford the investigation of the proposed research questions by the analysis of citation impact and knowledge dissemination within LIS. The choice of internal citations is consistent with other network-based citation analysis (e.g., Leydesdorff, 2007; Ma, Guan, & Zhao, 2008; Yan & Sugimoto, 2011). These citation instances were then aggregated into the identified 50 topics based on θ_d. The topic citation impact $CI(t(j), t)$ for topic j at year t is:

$$CI(t(j),t)=\sum_{d|py(d)=t}\theta_{d,t(j)}\times cr(d)$$

where $CI(t(j), t)$ denotes citation impact of topic j at year t and $cr(d)$ denotes the number of citations document d received. This fractional counting aligns with the citation impact assessment for authors (e.g., Egghe, 2008), journals (e.g., Zitt & Small, 2008), and institutions (e.g., Leydesdorff & Shin, 2011). For an effective comparison of topic citation impact across different years, $CI(t(j), t)$ was normalized by the sum of citations of all topics for year t (i.e., the total number of citations received by papers published in year t):

$$N\_CI(t(j),t)=\frac{\sum_{d|py(d)=t}\theta_{d,t(j)}\times cr(d)}{\sum_{j}\sum_{d|py(d)=t}\theta_{d,t(j)}\times cr(d)}$$

Using the $N\_CI(t(j), t)$ as the dependent variable and year of publication as the independent variable, a single-factor linear regression model can be formed. The slope of the modeled regression curve was used to measure the trend of topic impact.

Let PCN be the 47,137 by 47,137 paper citation network, whereas θ is the per-document topic proportion for all documents that has a dimension of 47,137 by 50. The topic-to-topic citation network (TCN) can thus be obtained by matrix manipulation $\theta' \times PCN \times \theta$. The dimension of TCN is 50 by 50. Using the same example topic proportion in Figure 3 and a three-paper PCN, Figure 4 shows the way to form a topic citation network.

As with disciplines, topics may exhibit different citation behaviors: Some topics may be more permeable and "inter-topical," whereas others may be more self-contained. Topic permeability was measured through topic self-citations:

$$Per(t(j))=\frac{TCN_{jj}}{\sum_{i}TCN_{ij}}$$

where $Per(t(j))$ denotes permeability of topic $t(j)$, TCN_{jj} denotes self-citation of topic $t(j)$, and TCN_{jj} denotes citation from topic $t(i)$ to topic $t(j)$.

Currently, clustering and mapping techniques are largely designed for co-occurrence-based networks (e.g., White & McCain, 1998; White, 2003; Rafols, Porter, & Leydesdorff, 2010). A different mapping strategy is needed to effectively

DOI: 10.1002/asi

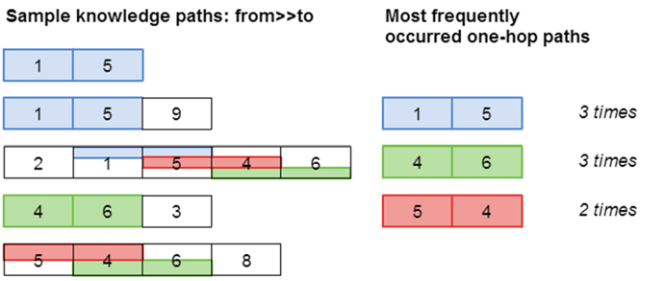

FIG. 5. An example of finding the most frequently occurring one-hop paths. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

represent a dense, small-scaled, directed network. In this study, shortest path was employed as the instrument to map knowledge flow among topics. Shortest paths characterize the most important knowledge paths in the topic citation network. Shortest path is typically defined in a distance-based network. The topic-to-topic citation network was thus transformed into a distance-based network through $reverse\_flow\_width_{i\to j} = \frac{1}{number\ of\ citations\ from\ j\ to\ i}$. It shows that the more citations from one topic to another, the wider the knowledge flow (Yan, 2014b). All 2,500 paths (i.e., 50*50 by including paths to themselves) were identified among the 50 topics through the Dijkstra algorithm. The most frequently occurred one-hop paths (i.e., paths that only involve two nodes) were extracted (see an example in Figure 5). They were used as the backbone of topical knowledge diffusion.

Results

Topics in Library and Information Science

The 50 topics were labeled by the top five words that have the highest associations with each topic, as seen in Table 1. For each topic, average year of publication, peak year, popularity and slope, and citation impact and slope are also included. Because earlier years may not provide sufficient data points for the linear regression model, slopes for topic popularity were calculated based on publications between 1964 and 2013. Because papers published in 2013 were cited very sparsely because of a narrow citation window, slopes for topic impact were calculated based on publications between 1964 and 2012.

Based on average year of publication, 23 topics belong to the 2000s, 25 to the 1990s, and three to the 1980s. The three most recent topics are t42 on online social networks, t24 on online technologies, and t26 on health communications. The three most dated topics are t17 on books and literature, t14 on cataloging, and t21 on history of legal services. Popularity of the 50 topics ranges from 604 to 1,296 with a mean of 942. The three most popular topics are t4 on scientific collaboration, t23 on information retrieval, and t24 on online technologies. Measured by slope of popularity, 34 topics have positive slopes and thus have gained popularity; in the meantime, 16 topics have negative slopes and are thus becoming somewhat less visible in LIS. Citation impact of the 50 topics ranges from 560 to 5,262 with a mean of 1,879. The top three topics that have the highest topic impact are t1 on users and technology, t4 on scientific collaboration, and t16 on citation analysis. Measured by slope of citation impact, 32 topics have positive slopes and 18 have negative slopes.

Topic Popularity

The dynamic aspect of topic popularity is assessed through $N\_Pop(t(j), t)$. In Figure 6, x-axis denotes year of publication (1964–2013) and y-axis denotes $N\_Pop(t(j), t)$ for the top five topics that have the steepest popularity loss (left panel) and popularity gain (right panel), measured by slope.

In Figure 7, while topics on cataloging, collection, and librarianship have experienced the most noticeable popularity decrease, topics on online technologies, data and knowledge management, and health communications have the most evident popularity gain.

Topic Impact

The dynamic aspect of topic impact is measured through $N\_CI(t(j), t)$. In Figure 7, the x-axis denotes year of publication (1964–2012) and the y-axis denotes $N\_CI(t(j), t)$ for the top five topics that have the steepest impact loss (left panel) and impact gain (right panel), measured by slope.

Compared with dynamics of topic popularity, the dynamics of topic impact is subject to higher degrees of variation. This is attributed to the fact that preeminent papers can receive up to several hundred citations, causing yearly topic impact fluctuations. Topics on literature, classification, and document retrieval were highly cited in the 1960s and 1970s; however, the concentration was shifted to online technologies, knowledge management, and journal citation impact analysis thereafter.

Discussion

Statistical and Qualitative Evaluations of the Results

JSD was employed to evaluate the quality of the 50 topics. JSD is calculated based on word topic assignment. The lower bound of JSD is 0, the situation that two word topic assignments are identical; the upper bound is 0.7 (In[2]), the situation that two topics are completely different. Figure 8 shows the heatmap of the topic dissimilarity matrix measured by JSD.

Figure 8 illustrates that most topics have high JSD scores with respect to other topics. This is a good sign that the topic model has successfully identified distinctive topics. We also see that some topics are topically related to others. This is expected because different topics may share a (small)

DOI: 10.1002/asi

TABLE 1. Features of the 50 topics in Library and Information Science.

ID	Topic	Average year of publication	Peak year	Popularity	Slope (1964–2013)	Citation impact	Slope (1964–2012)
t1	factors-user-technology-use-effects	2001.69	2011	1,103.09	5.19E-04	5,261.53	1.06E-03
t2	papers-hot-biology-human-cell	1998.14	1993	613.08	2.54E-04	559.89	1.02E-04
t3	science-review-technology-issues-international	1992.80	1969	901.94	−8.10E-04	2,144.57	−1.71E-04
t4	scientific-science-collaboration-international-bibliometric	2000.54	2013	1,296.55	5.00E-04	4,812.41	7.28E-04
t5	document-review-literature-supply-delivery	1996.89	1987	678.61	−2.28E-05	955.30	1.23E-04
t6	software-development-systems-source-management	2002.15	2013	922.71	4.77E-04	2,239.13	7.64E-04
t7	law-model-function-data-new	1997.83	2013	797.12	1.20E-05	2,670.05	−1.65E-03
t8	data-using-spatial-gis-model	2003.21	2000	1,018.78	5.37E-04	1,578.77	5.45E-04
t9	government-public-national-federal-policy	1992.21	1966	1,002.58	−7.50E-04	903.35	−2.19E-04
t10	model-theory-models-seeking-design	2000.84	2007	916.93	3.27E-04	2,500.87	6.57E-04
t11	systems-support-decision-system-design	1995.48	1987	998.52	−5.11E-05	1,415.88	9.19E-05
t12	collections-collection-development-special-problems	1991.39	1965	893.09	−1.11E-03	991.43	−6.51E-04
t13	paper-conservation-effect-treatment-influence	1996.32	1971	604.48	−9.11E-05	1,045.20	6.68E-05
t14	cataloging-collection-book-circulation-bibliographic	1986.85	1965	967.06	−1.89E-03	1,030.31	−7.27E-04
t15	reference-service-services-virtual-academic	1997.69	1965	789.31	−8.15E-05	1,553.20	1.16E-04
t16	citation-impact-journal-science-bibliometric	2001.30	2013	973.37	4.27E-04	4,811.42	7.34E-04
t17	books-book-literature-work-years	1984.75	1966	1,002.28	−1.73E-03	721.69	−3.64E-04
t18	access-electronic-publishing-digital-scholarly	2001.01	2000	1,007.33	3.84E-04	1,332.99	2.19E-04
t19	theory-making-communication-systems-practice	2001.15	2013	877.85	3.67E-04	1,758.16	3.00E-04
t20	health-informatics-medical-care-sciences	2001.88	2002	900.54	2.81E-04	1,182.73	3.47E-04
t21	law-legal-american-history-public	1988.46	1968	892.53	−1.52E-03	725.72	−1.00E-03
t22	classification-indexing-using-text-automatic	1995.68	1971	959.63	−4.07E-04	1,140.08	−1.66E-03
t23	retrieval-using-query-relevance-document	1998.09	1971	1,241.10	2.31E-05	1,221.65	−1.50E-03
t24	empirical-online-technology-electronic-model	2004.23	2013	1,143.17	7.92E-04	4,033.81	1.38E-03
t25	online-searching-catalog-subject-bibliographic	1990.33	1977	1,071.41	−9.59E-04	1,816.81	−1.24E-03
t26	health-cancer-communication-internet-among	2004.00	2006	976.63	5.94E-04	1,318.15	5.05E-04
t27	language-medical-using-knowledge-data	2000.78	1998	1,060.60	3.81E-04	776.74	1.00E-05
t28	use-university-electronic-students-resources	2000.05	1968	903.33	1.10E-04	2,796.27	−6.85E-04
t29	scientists-science-researchers-new-with	1995.82	1990	862.59	9.03E-05	1,757.27	1.18E-04
t30	more-century-better-people-work	1996.97	1965	684.85	6.16E-06	959.11	7.82E-05
t31	copyright-privacy-policy-issues-security	1997.55	1968	806.68	1.84E-05	895.64	2.11E-04
t32	knowledge-management-organizational-technology-innovation	2003.83	2011	1,032.81	6.22E-04	2,344.99	9.06E-04
t33	clinical-electronic-care-system-patient	2002.42	1999	1,067.18	6.20E-04	598.62	2.61E-04
t34	impact-journals-citation-journal-scientific	2000.90	2013	971.52	3.94E-04	3,540.85	6.46E-04
t35	web-sites-world-site-wide	2003.25	2004	846.21	5.36E-04	1,985.14	3.80E-04
t36	learning-literacy-students-education-instruction	2001.11	2011	1,017.50	2.66E-04	1,953.58	5.42E-04
t37	management-systems-information-systems-issues-implementation	1998.38	1994	963.11	1.65E-04	2,499.09	1.81E-04
t38	literature-citation-science-scientific-networks	1998.75	1974	1,019.32	7.39E-05	3,733.02	−1.83E-03
t39	quality-service-university-evaluation-interlibrary	1996.71	1966	858.04	−2.73E-04	1,707.26	−7.23E-04
t40	telecommunications-mobile-market-policy-competition	1999.70	1992	1,115.86	4.65E-04	1,712.36	4.03E-04
t41	search-web-searching-user-users	2001.55	2007	888.70	4.42E-04	1,870.88	−1.25E-04
t42	social-online-network-communities-networks	2004.77	2012	907.75	5.36E-04	1,273.15	4.68E-04
t43	years-science-next-new-society	1996.60	1987	789.17	−1.17E-04	1,020.69	1.58E-04
t44	business-technology-value-process-small	1999.95	1995	884.66	3.41E-04	2,405.64	5.72E-04
t45	know-good-question-online-find	1997.54	1965	1,044.31	1.31E-04	1,955.48	2.22E-04
t46	new-education-future-librarianship-technology	1991.34	1971	932.57	−9.66E-04	1,236.03	−2.40E-04
t47	data-digital-system-database-metadata	2000.83	1997	1,066.95	4.53E-04	1,206.70	7.64E-05
t48	librarians-academic-librarian-faculty-professional	1992.56	1968	1,046.59	−7.17E-04	2,112.04	−7.26E-04
t49	digital-developing-africa-south-countries	2002.60	2011	878.81	3.62E-04	1,248.06	4.69E-04
t50	science-knowledge-theory-social-systems	1998.30	1969	938.24	−6.83E-06	2,617.24	6.17E-05

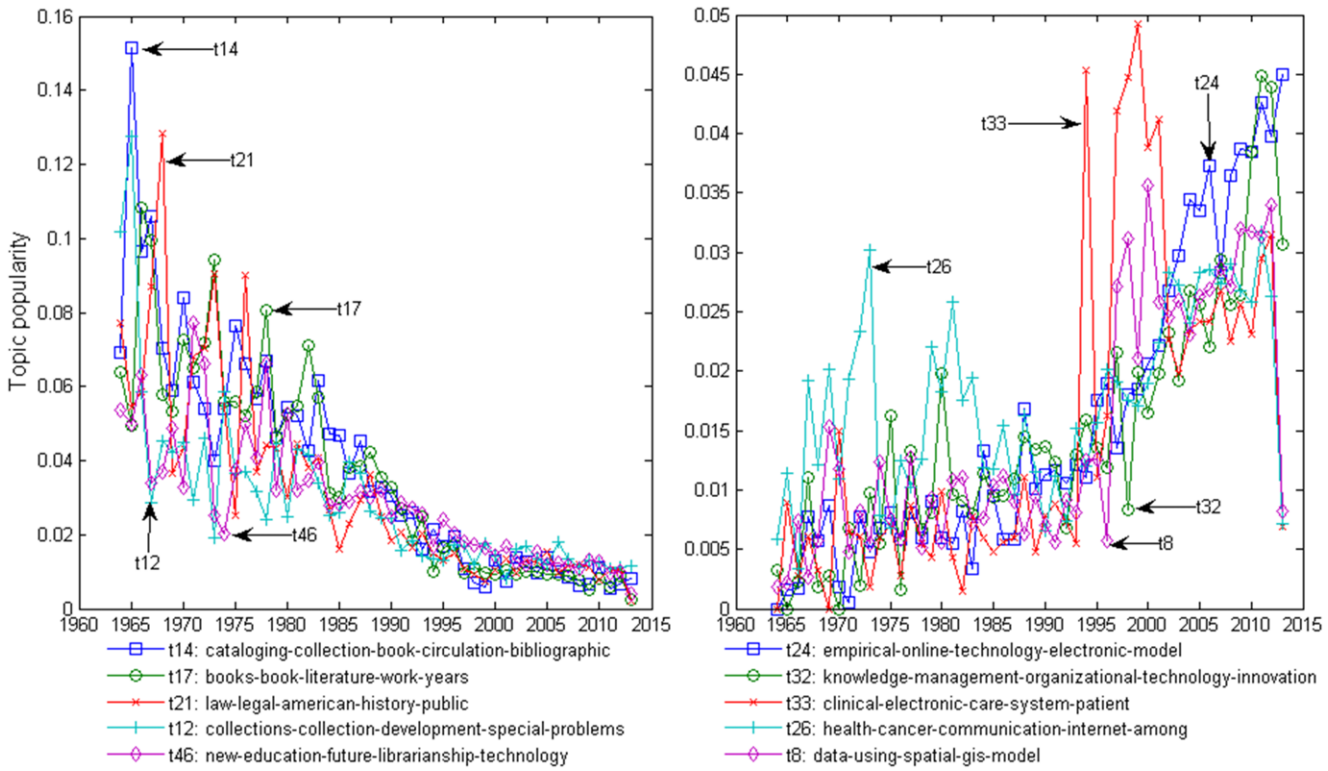

FIG. 6. Topics that have the steepest popularity loss (left panel) and popularity gain (right panel). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

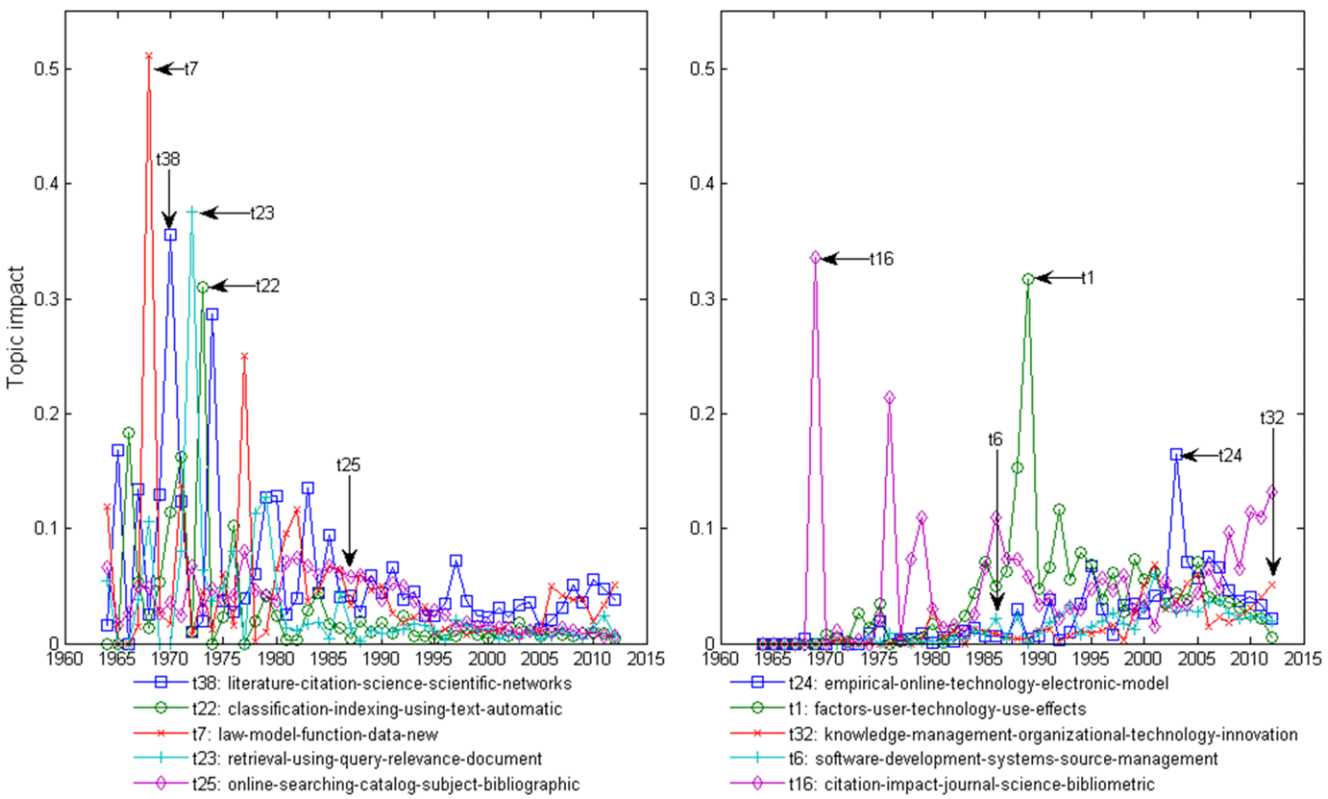

FIG. 7. Topics that have the steepest impact loss (left panel) and impact gain (right panel). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

DOI: 10.1002/asi

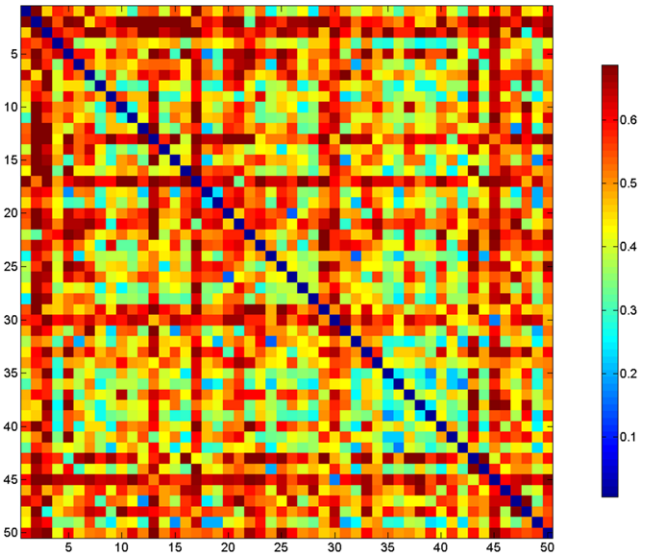

FIG. 8. A heatmap presentation of JSD for the 50 topics. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

collection of common lexicons (Blei & Lafferty, 2007); this is especially true for topics from a single field, such as LIS. While the use of JSD is able to establish some statistical certainty, the real-world implications should also be examined. Thus, we compare the results of this study with previous analyses of the cognitive structure of LIS and use this as an opportunity to evaluate the empirical implications.

One category of analysis has primarily focused on the identification of the epistemological divisions of LIS. It has been argued that information science and library science have philosophical and theoretical differences (e.g., Brookes, 1980; Hjørland, 2000; Ma, 2012). Quantitative studies have supported such an argument; for instance, through cocitation and bibliographic coupling analysis, studies have identified subfields in LIS: library science; information science and technology (information retrieval); and informetrics (Åström, 2002; Milojević, Sugimoto, Yan, & Ding, 2011; Waltman, Yan, & Van Eck, 2011). Results obtained from this study align with the previous analyses in that the 50 topics can be grouped into these three subfields.

In addition to top-level divisions, studies have also identified research clusters in LIS. The pioneering author cocitation analysis by White and Griffith (1981) used factor analysis to identify five clusters between 1972 and 1975, including scientific communication, bibliometrics, generalists, information retrieval (IR), and precursors of LIS. Author cocitation analysis was later used by White and McCain (1998) to identify 12 research clusters between 1972 and 1995, including experimental retrieval, citation analysis, online retrieval, bibliometrics, general library systems, science communication, user theory, online public access catalogs, imported ideas, indexing theory, citation theory, and communication theory. In a follow-up article by White (2003), a novel pathfinder network was used to visualize author cocitation relations and the same clusters were revealed (with the exception that citation theory was no longer a stand-alone cluster). A recent author cocitation analysis by Zhao and Strotmann (2014) has suggested a different set of clusters between 2006 and 2010. These include information behavior, bibliometric distributions, mapping of science, relevance, IR systems, webometrics, bibliometrics and science and innovation systems, use of e-resources, information systems (IS) theory and foundation, knowledge management, and text categorization. Results obtained from this study are consistent with these previous efforts. For instance, it has delineated topics broadly related to informetrics (t4, t16, and t34), IR (t23, t25, and t41), communication theory (t19), knowledge management (t32 and t37), indexing theory (t22), user theory (t1), information behavior (t10), and use of e-resources (t28). In the meantime, we also identified topics that were not included by previous studies, such as topics on health communication (t26), health informatics (t26), electronic clinical record (t33), social networks (t42), software development (t6), and IS (t11). This study has also identified precursory topics of some of these topics (see section on Knowledge Flow and Inheritance).

By matching the results of this study with previous diachronical analysis, some consensus can be observed. This study found that topics on users, bibliometrics, online technologies, communications, health informatics, and knowledge management are becoming more popular in LIS. It confirms Larivière, Sugimoto, and Cronin's (2012) study that words such as citation, impact, bibliometrics, use, management, health, clinical, and network became popular title words in LIS between 1900 and 2010. A study by Milojević et al. (2011) also found that title terms such as citation, impact factor, and web increased between 1989 and 2008. Topics such as collections, books and literature, indexing, and librarianship are becoming less popular. The result is consistent with Larivière et al.'s (2012) study.

The set of approaches used in this study has three advantages. First, the topic modeling technique works directly with words and does not require domain knowledge to interpret the clusters of papers, authors, or journals. This technique relates to co-word analysis (e.g., Janssens et al., 2008; Milojević et al., 2011; Yan et al., 2012). Different from co-word analysis, the topic modeling technique assigns words to clusters based on certain probabilistic distributions. Co-word analysis, on the other hand, typically partitions words to mutually exclusive clusters. Second, the number of topics identified in this study is considerably larger than was obtained from co-occurrence-based network analyses. The sizable number of topics grants a more granular analysis of the cognitive structure of LIS. In addition to probing into the top-level divisions or major clusters of LIS, this study reveals a more detailed research landscape. Such a level of analysis is usually not attainable through co-occurrence-based network

DOI: 10.1002/asi

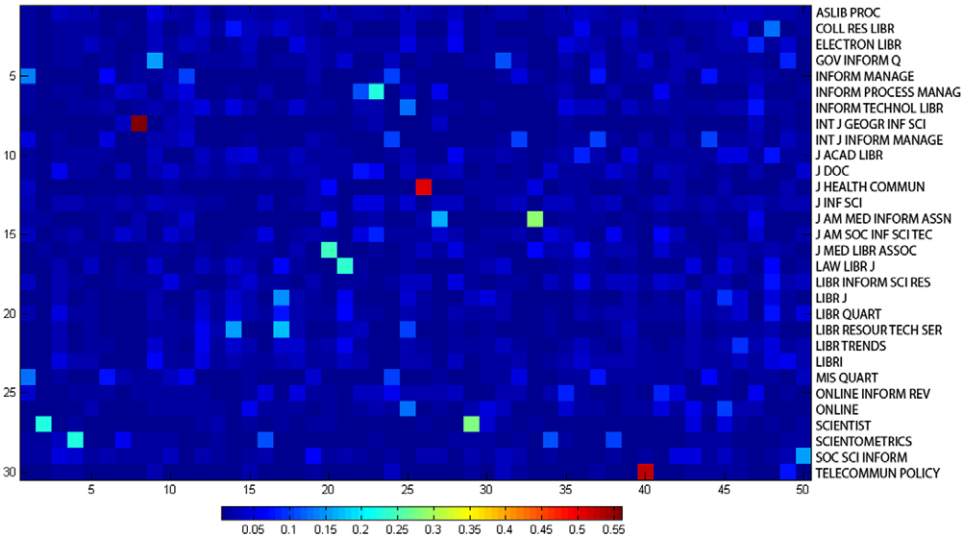

FIG. 9. Heatmap presentation of journal topical specializations. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

analysis because of the complexity of some densely connected co-occurrence clusters. Finally, the proposed dynamic analysis method has streamlined the diachronical analysis of topic popularity and impact. The rendering of yearly topic popularity is an improvement over time-sliced methods. Furthermore, the integration of citation information with topic analysis provides an opportunity to examine the intertopic knowledge diffusion patterns in LIS. These advances complement co-occurrence-based network studies.

Journal-Level Topical Diversity

To quantitatively measure journal-level topical diversity, we use a heatmap (Figure 9) to show the distribution of topics on the most productive 30 journals in LIS. Cell values in the heatmap are calculated through aggregating θ_d for each journal under each topic and then normalizing by the total number of topic proportions for each journal. The journal-level topical diversity of journal $j(s)$ for topic j can thus be expressed as $JTD_{j(s)} = \frac{\sum_{d|d\in j(s)} \theta_{d,t(j)}}{\sum_j \sum_{d|d\in j(s)} \theta_{d,t(j)}}$. Cell values range from 0 to 1, with 0 being no specialization in the topic and 1 being exclusively specialized in the topic.

Although a handful of journals (e.g., *International Journal of Geographical Information Science*, *Journal of Health Communication*, *Journal of the American Medical Informatics Association*, and *Telecommunications Policy*) have more-specialized foci, the majority of journals have wider coverage. The result suggests that these journals do not necessarily publish papers solely on one research specialty, but contribute to an array of specialties. These journals may represent LIS as a field of study, but they may not have the granularity to signify the different research specialties within LIS.

Shannon entropy (Figure 10) is adopted to illustrate the dynamic changes of journal topical diversity. It measures, for each journal, the proportions of its papers under each topic:

$$H = -\sum_{j=1}^{50} \left(JTD_{j(s)} \ln JTD_{j(s)} \right) = -\sum_{j=1}^{50} \frac{\sum_{d|d\in j(s)} \theta_{d,t(j)}}{\sum_j \sum_{d|d\in j(s)} \theta_{d,t(j)}} \ln \frac{\sum_{d|d\in j(s)} \theta_{d,t(j)}}{\sum_j \sum_{d|d\in j(s)} \theta_{d,t(j)}}$$

for journal s. Similar to JSD, Shannon entropy is also a measure of dissimilarity in that the higher the value, the more diverse the content; different from JSD, though, it measures the congruence of a single object based on its variables, whereas JSD is typically applied to measure the dissimilarity of two objects.

Diachronically, most journals in Figure 10 have either a steady or increased level of diversity, with the exception of two health informatics journals (*Journal of the American Medical Informatics Association* and *Journal of the Medical Library Association*) and two professional journals (*Library Journal* and *Online*). A few journals have a steeper diversity gain, such as *Government Information Quarterly*, *Journal of Documentation*, and *Library & Information Science Research*. It indicates that these journals are embracing a broader research scope and publishing papers that cover more extensive topics. Because journals are becoming more intertopical and even interdisciplinary, simply relying on journal-level analyses to examine the substrate of various research fields may be less effective. Based on this evidence

DOI: 10.1002/asi

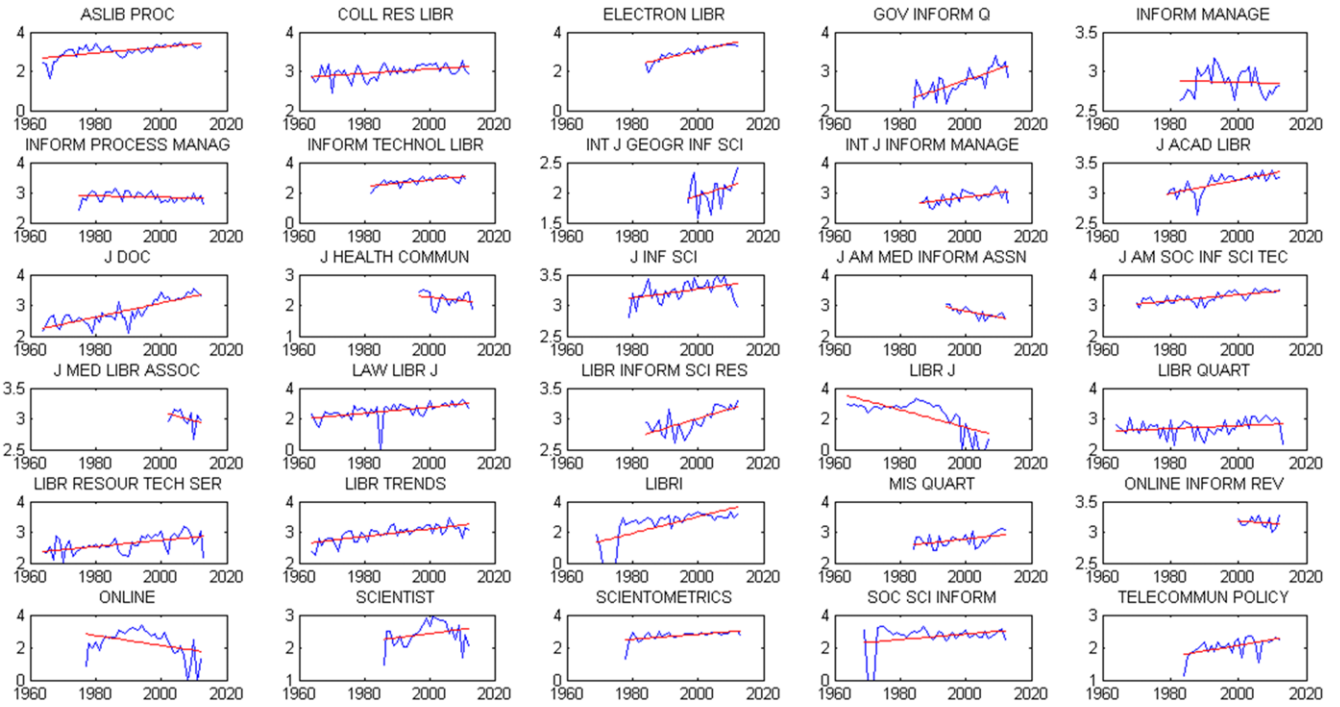

FIG. 10. Shannon entropy for the most productive 30 journals in LIS. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

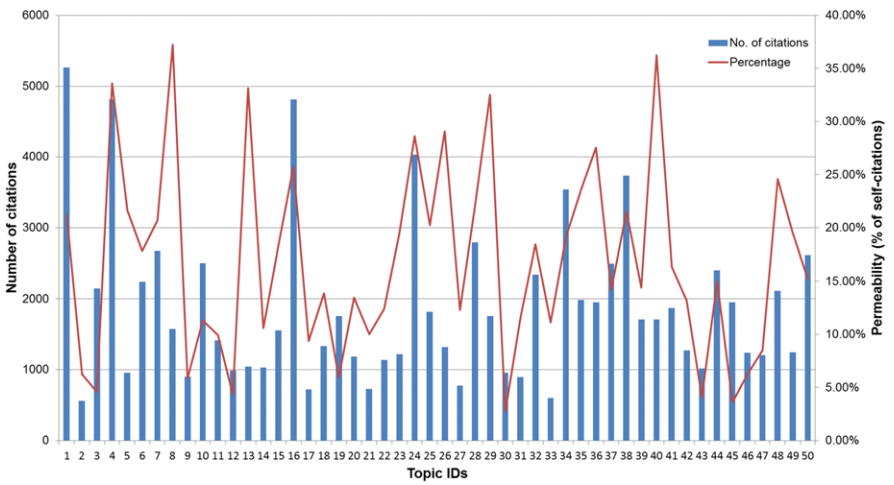

FIG. 11. Topic self-dependence (blue bars: number of citations; red lines: percentage of self-citations). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

therefore, we argue that topic-level studies are needed to take analysis to a newer level of granularity.

Topic Popularity Versus Impact

Similar to authors and journals, topics also possess multiple attributes. In addition to social and cognitive attributes (e.g., Yan, Ding, Milojević, & Sugimoto, 2012), this study focuses on two other attributes: popularity and impact. It first examines topic dependence though self-citations (Figure 11), then the coevolving feature of topic popularity and impact between 1964 and 2012 (Figure 12), and, finally, the correlation relationships among topic popularity, impact, and year of publication (Figure 13).

In Figure 11, one topic has a citation impact of more than 5,000, five between 3,000 and 5,000, 34 between 1,000 and 3,000, and 10 below 1,000. Topics that have the highest citation impact are t1 on users and technology, t4 on scientific collaboration, t16 on citation analysis, t24 on online technologies, and t38 on science literature. Highly cited

DOI: 10.1002/asi

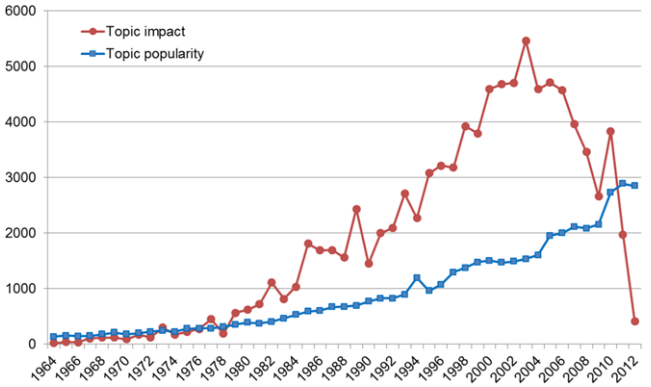

FIG. 12. Accumulated topic popularity and impact between 1964 and 2012. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

topics are situated in two broad research areas: informetrics and information systems and technologies.

Similar to disciplines (Yan, Ding, Cronin, & Leydesdorff, 2013), research topics also vary in terms of the degree of permeability and can be quantified through topic self-citation ratios. The results show that five topics have self-citation ratios between 0.3 and 0.4, 12 between 0.2 and 0.3, 21 between 0.1 and 0.2, and 12 below 0.1. The most self-contained topics include t8 on geographic information systems (GIS) and spatial models, t40 on telecommunication, and t4 on scientific collaboration. These topics have a more distinctive core, and their papers are published on more-specialized subject matters.

In Figure 12, while topic popularity has increased steadily, topic impact reached its peak in 2003 and decreased thereafter. Because the data set was collected in 2013, the 9-year window indicates that publications in the field of LIS as a whole reached their highest impact 9 years after publication. The width of the window is dependent on the number of citable publications for each year and the immediacy of publications being cited.

Because of the nonparametric distribution pattern of citation impact, popularity, and average year of publication, the Spearman rank correlation coefficient was used to test the relationships between these variables. The result shows that only the correlation between impact and average year of publication is statistically significant ($r = 0.372$, $p < .05$). This indicates that newer topics tend to be cited more often than older topics. This may be attributed to the fact that there has been a steady increase of citable publications in recent years. The correlation between impact and popularity is not statistically significant ($r = 0.279$, $p = .056$), which suggests a nonsignificant overlap between topic popularity and topic impact. Such a nonsignificant relationship also validates the need to have both topic popularity and topic impact in describing topic characteristics. The correlation coefficient between popularity and average year of publication is 0.193 ($p = .18$), which suggests a nonsignificant relationship between the age of topics and their likelihood of gaining popularity.

Topical Knowledge Flow and Inheritance

In this subsection, we examine topical knowledge dissemination using shortest paths. In Figure 14, the most frequently occurred paths are illustrated. Each node represents one topic and is color coded to reflect its average year of publication. The size of a node denotes the number of times a topic occurred on the shortest paths. In total, 511 one-hop knowledge paths are identified. The top 50 paths based on occurrence frequency constitute the primary knowledge path in Figure 14. The frequency ranges from 203 to 29. Eighteen topics are not included by these 50 paths. They are connected through secondary knowledge paths.

Several knowledge hubs can be identified: t1 on users and technology, t4 on scientific collaboration, t16 on citation analysis, t24 on online technologies, and t38 on science literature play an important role in transferring and dissemination topical knowledge. Three major topical knowledge paths are:

- From electronic resources to digital data and information systems, then to information retrieval systems, then to telecommunication systems, online social networks, health communication, and knowledge management;
- From journal and paper citation impact analysis to collaboration and bibliometrics, then to web science; and
- From collection, literature, and reference to health informatics and GIS.

Older topics are largely connected by secondary knowledge paths. This suggests that these topics are no longer active in importing or exporting topical knowledge. They are gradually fading away from the research frontier of LIS. In the meantime, newer topics are tightly connected with the rest of the knowledge flow network. Thus, their topic-level knowledge is expected be incorporated into the knowledge base of LIS in the future.

Conclusion

This study conducted a topic-level analysis using a data set on library and information science publications. It highlighted the use of topics in organizing scientific literature and examining research popularity, impact, and dynamics. It employed a topic modeling technique and conducted a series of data analytics, including dynamic analysis, regression, shortest path, correlation, and similarity analysis. It found that topics on online technologies, informetrics, IR systems, health communication and informatics, and online social networks have gained popularity over the past decades. Topics related to literature, books, collections, and cataloging, on the other hand, have declined in popularity. These findings therefore have addressed the first research question on the dynamic characteristics of topics in LIS. The study has also revealed that LIS journals are becoming more inter-topical. This finding suggests that topic-level studies are

DOI: 10.1002/asi

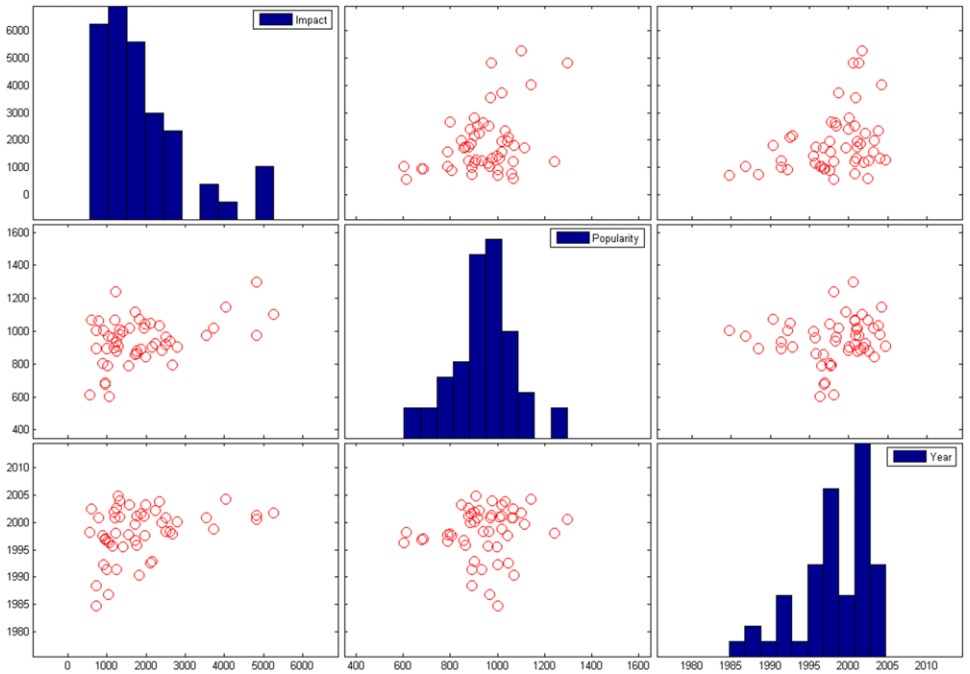

FIG. 13. Scatterplots and histograms of topic impact, popularity, and average year of publication. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

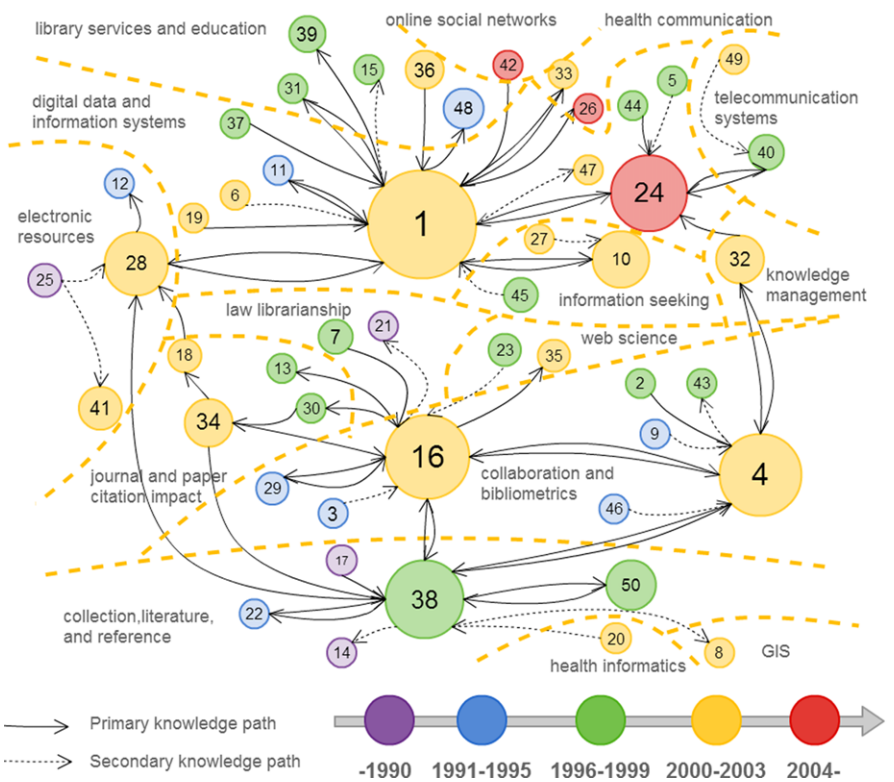

FIG. 14. Topical knowledge flow. Color coding: average year of publication; size of node: number of times a topic is located on the shortest paths; links: most frequently occurred one-hop paths. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

DOI: 10.1002/asi

necessary to bring more granular perspectives to existing co-occurrence-based analyses.

By adding citation information, topics with the highest impact have been identified, including those on technology use, collaboration and bibliometrics, citation analysis, and online technologies. This study also found several topics that are more self-contained than others, including those on GIS and spatial models, telecommunication, scientific collaboration, and health communication. Through a correlation analysis, this study verified a nonsignificant relationship between topic popularity and topic impact. The result helped answer the second research question and also validates the need to have both attributes to describe topic characteristics.

Furthermore, a topical knowledge flow network revealed topic-level knowledge dissemination channels. The study identified a few major knowledge paths in LIS, including those from electronic resources to online social networks and knowledge management, from journal and paper citation impact to web science, and from collection, literature, and reference to health informatics and GIS. It also found that whereas older topics are becoming less active in exporting and importing knowledge, newer topics are becoming integral components in the LIS knowledge flow network. These findings address the last research question on knowledge dissemination patterns of topics in LIS.

This paper explored topic dynamics and discovered patterns of topic impact and popularity. In order to determine factors that contribute to such dynamic changes, one needs to relate textual data with other data sources (e.g., authorship data and funding information). Future studies will benefit from identifying the mechanism (e.g., shifts in research communities and science policies) that leads to changes in topic impact and popularity.

Acknowledgments

We would like to thank David McAllister and Yongjun Zhu for their comments to an earlier version of this paper.

References

Åström, F. (2002). Visualizing Library and Information Science concept spaces through keyword and citation based maps and clusters. In The Fourth International Conference on Conceptions of Library and Information Science (CoLIS4) (pp. 185–197). Greenwood Village, CO: Libraries Unlimited.

Åström, F. (2007). Changes in the LIS research front: Time-sliced cocitation analyses of LIS journal articles, 1990–2004. Journal of the American Society for Information Science and Technology, 58(7), 947–957.

Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., & Weitzner, D. (2006). Creating a science of the Web. Science, 313(5788), 769–771.

Björk, B.C., Welling, P., Laakso, M., Majlender, P., Hedlund, T., & Guðnason, G. (2010). Open access to the scientific journal literature: Situation 2009. PLoS ONE, 5(6), e11273.

Blei, D.M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77–84.

Blei, D.M., & Lafferty, J.D. (2007). A correlated topic model of science. The Annals of Applied Statistics, 1(1), 17–35.

Blei, D.M., & Lafferty, J.D. (2009). Topic models. Retrieved from http://www.cs.princeton.edu/~blei/papers/BleiLafferty2009.pdf

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3(4–5), 993–1033.

Bolelli, L., Ertekin, S., Zhou, D., & Giles, C.L. (2009). Finding topic trends in digital libraries. In Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital libraries (pp. 69–72). New York: ACM.

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

Borgman, C.L., & Rice, R.E. (1992). The convergence of information science and communication: A bibliometric analysis. Journal of the American Society for Information Science, 43(6), 397–411.

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, R. (2014). Creation of a highly detailed, dynamic, global model and map of science. Journal of the Association for Information Science and Technology, 65(4), 670–685.

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

Brookes, B.C. (1980). The foundations of information science. Part I. Philosophical aspects. Journal of Information Science, 2(3–4), 125–133.

Chen, C.M. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl. 1), 5303–5310.

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.

Cohn, D., & Hofmann, T. (2000). The missing link: A probabilistic model of document content and hypertext connectivity. In Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference (Vol. 13, pp. 430–437). Cambridge, MA: The MIT Press.

Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. In Proceedings of the 24th International Conference on Machine Learning (pp. 233–240). New York: ACM.

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.

Egghe, L. (2008). Mathematical theory of the h- and g-index in case of fractional counting of authorship. Journal of the American Society for Information Science and Technology, 59(10), 1608–1616.

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed-membership models of scientific publications. Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl. 1), 5220–5227.

Evans, J.A., & Reimer, J. (2009). Open access and global participation in science. Science, 323(5917), 1025–1025.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. Scientometrics, 56(3), 357–367.

Griffiths, T.L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl. 1), 5228–5235.

Hall, D., Jurafsky, D., & Manning, C.D. (2008). Studying the history of ideas using topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 363–371). Stroudsburg, PA: Association for Computational Linguistics.

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, L. (2009). Detecting topic evolution in scientific literature: How can citations help? In Proceedings of the 18th ACM Conference on Information and Knowledge Management (pp. 957–966). New York: ACM.

Hjørland, B. (2000). Library and information science: Practice, theory, and philosophical basis. Information Processing and Management, 36(3), 501–531.

Holton, G. (1978). Scientific imaginations: Case studies. Cambridge, UK: Cambridge University Press.

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

DOI: 10.1002/asi

Järvelin, K., & Vakkari, P. (1993). The evolution of library and information science 1965–1985: A content analysis of journal articles. Information Processing and Management, 29(1), 129–144.

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

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.

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

Leydesdorff, L., & Probst, C. (2009). The delineation of an interdisciplinary specialty in terms of a journal set: The case of communication studies. Journal of the American Society for Information Science and Technology, 60(8), 1709–1718.

Leydesdorff, L., & Shin, J.C. (2011). How to evaluate universities in terms of their relative citation impacts: Fractional counting of citations and the normalization of differences among disciplines. Journal of the American Society for Information Science and Technology, 62(6), 1146–1155.

Leydesdorff, L., & Vaughan, L. (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment. Journal of the American Society for Information Science and Technology, 57(12), 1616–1628.

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.

Ma, L. (2012). Meanings of information: The assumptions and research consequences of three foundational LIS theories. Journal of the American Society for Information Science and Technology, 63(4), 716–723.

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

Masada, T., & Takasu, A. (2012). Extraction of topic evolutions from references in scientific articles and its GPU acceleration. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management (pp. 1522–1526). New York: ACM.

Metzger, N., & Zare, R.N. (1999). Interdisciplinary research: From belief to reality. Science, 283(5402), 642–643.

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.

Nallapati, R.M., Ahmed, A., Xing, E.P., & Cohen, W.W. (2008). Joint latent topic models for text and citations. In Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 542–550). New York: ACM.

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

Prebor, G. (2010). Analysis of the interdisciplinary nature of library and information science. Journal of Librarianship and Information Science, 42(4), 256–267.

Rafols, I., & Leydesdorff, L. (2009). Content-based and algorithmic classifications of journals: Perspectives on the dynamics of scientific communication and indexer effects. Journal of the American Society for Information Science and Technology, 60(9), 1823–1835.

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.

Ramage, D., Hall, D., Nallapati, R., & Manning, C.D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (pp. 248–256). Stroudsburg, PA: Association for Computational Linguistics.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (pp. 487–494). Arlington, Virginia: AUAI Press.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science, 24(4), 265–269.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. Handbook of Latent Semantic Analysis, 427(7), 424–440.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 306–315). New York: ACM.

Sugimoto, C.R., Li, D., Russell, T.G., Finlay, S.C., & Ding, Y. (2011). The shifting sands of disciplinary development: Analyzing North American Library and Information Science dissertations using latent Dirichlet allocation. Journal of the American Society for Information Science and Technology, 62(1), 185–204.

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 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 990–998). New York: ACM.

Tu, Y., Johri, N., Roth, D., & Hockenmaier, J. (2010). Citation author topic model in expert search. In Proceedings of the 23rd International Conference on Computational Linguistics (pp. 1265–1273). Stroudsburg, PA: Association for Computational Linguistics.

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

Van Raan, A.F.J. (2004). Measuring science: Capita Selecta of current issues. In H.F. Moed, W. Glänzel, & U. Schmoch (Eds.), Handbook of quantitative science and technology research (pp. 19–50). Dordrecht, The Netherlands: Kluwer Academic.

Waltman, L., & Van Eck, N.J. (2012). A new methodology for constructing a publication-level classification system of science. Journal of the American Society for Information Science and Technology, 63(12), 2378–2392.

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.

Wang, C., & Blei, D.M. (2011). Collaborative topic modeling for recommending scientific articles. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 448–456). New York: ACM.

Wang, X., & McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 424–433). New York: ACM.

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

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. (2014a). Research dynamics: Measuring the continuity and popularity of research topics. Journal of Informetrics, 8(1), 98–110.

Yan, E. (2014b). Finding knowledge paths among scientific disciplines. Journal of the American Society for Information Science and Technology, 65(1), 2331–2347.

Yan, E., & Sugimoto, C.R. (2011). Institutional interactions: Exploring 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., Milojević, S., & Sugimoto, C.R. (2012). Topics in dynamic research communities: An exploratory study for the field of information retrieval. Journal of Informetrics, 6(1), 140–153.

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.

Yang, Z., Yin, D., & Davison, B.D. (2011). Award prediction with temporal citation network analysis. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1203–1204). New York: ACM.

Zhang, L., Liu, X., Janssens, F., Liang, L., & Glänzel, W. (2010). Subject clustering analysis based on ISI category classification. Journal of Informetrics, 4(2), 185–193.

Zhao, D., & Strotmann, A. (2014). The knowledge base and research front of information science 2006–2010: An author cocitation and bibliographic coupling analysis. Journal of the Association for Information Science and Technology, 65(5), 995–1006.

Zitt, M., & Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: The audience factor. Journal of the American Society for Information Science and Technology, 59(11), 1856–1860.

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.

DOI: 10.1002/asi