Journal of Informetrics 10 (2016) 606–621

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Generation of topic evolution trees from heterogeneous bibliographic networks

Scott Jensen [a], Xiaozhong Liu [b,*], Yingying Yu [c], Staša Milojevic [d]

[a] San Jose State University, Lucas College and Graduate School of Business, MIS Department, One Washington Square, San Jose, CA 95192, USA
[b] Indiana University, School of Informatics and Computing, Dept. of Info. and Library Sciences, Wells Library, Room 027, 1320 E. 10th Street, Bloomington, IN 47405, USA
[c] Dalian Maritime University, College of Transportation Management, Department of Management Science and Engineering, Room 211, Guanli Building, No. 1, Linghai Road, Dalian, Liaoning Province 116026, China
[d] Indiana University, School of Informatics and Computing, Dept. of Info. and Library Sciences, Wells Library, Room 017, 1320 E. 10th Street, Bloomington, IN 47405, USA

ARTICLE INFO

Article history:
Received 7 December 2015
Received in revised form 8 April 2016
Accepted 8 April 2016
Available online 6 May 2016

Keywords:
Topic evolution
Heterogeneous bibliographic network
Meta-path
Visualization

ABSTRACT

The volume of the existing research literature is such it can make it difficult to find highly relevant information and to develop an understanding of how a scientific topic has evolved. Prior research on topic evolution has often leveraged refinements to Latent Dirichlet Allocation (LDA) to identify emerging topics. However, such methods do not answer the question of which studies contributed to the evolution of a topic. In this paper we show that meta-paths over a heterogeneous bibliographic network (consisting of papers, authors and venues) can be used to identify the network elements that made the greatest contributions to a topic. In particular, by adding derived edges that capture the contribution of papers, authors, and venues to a topic (using PageRank algorithm), a restricted meta-path over the bibliographic network can be used to restrict the evolution of topics to the context of interest to a researcher. We use such restricted meta-paths to construct a topic evolution tree that can provide researchers with a web-based visualization of the evolution of a scientific topic in the context of interest to them. Compared to baseline networks without restrictions, we find that restricted networks provide more useful topic evolution trees.

1. Introduction

An exponential growth in the output of scientific literature (Bornmann & Mutz, 2015; Price, 1961, 1963; Van Raan, 2000) is one of the staples of contemporary science. In such an environment, scientists are becoming more specialized with narrower expertise (Jones, 2005, 2010), while the solutions of difficult problems in science and industry often require interdisciplinary approaches (Wagner et al., 2011) and team-based research (Falk-Krzesinski et al., 2010). Thus, ideally, scientists need to have deep knowledge in their own specialty and broad knowledge across a range of domains, i.e., be "T-shaped" (Barile, Franco, Nota, & Saviano, 2012; Donofrio, Spohrer, & Zadeh, 2010). While the Internet and electronic publications have made it easier to access an unprecedented volume and range of resources, this wealth of information can make it difficult to identify the best resources for learning the foundations of a specific topic or to identify the researchers, papers, and venues that have

* Corresponding author.
E-mail addresses: scott.jensen@sjsu.edu (S. Jensen), liu237@indiana.edu (X. Liu), uee870927@126.com (Y. Yu), smilojev@indiana.edu (S. Milojevic).

http://dx.doi.org/10.1016/j.joi.2016.04.002
1751-1577/

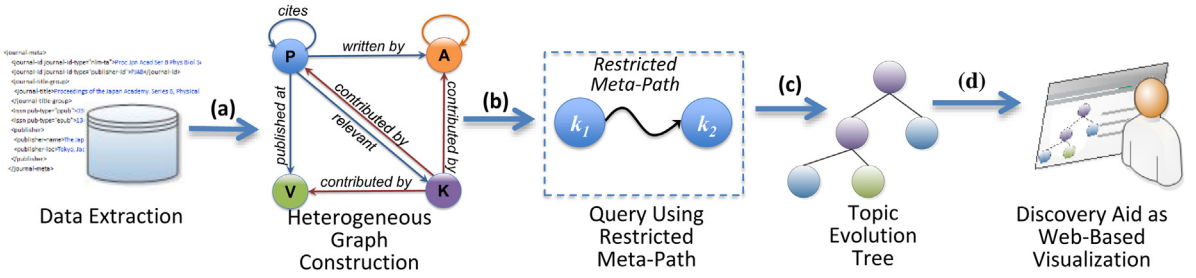

Fig. 1. Overview of the topic evolution tree (TET) methodology for identifying the evolution of scientific topics: (a) generating a heterogeneous graph from scholarly data (b) calculating the relatedness of topics in a user's context (c) generating a context-sensitive TET, and (d) visualizing the evolution of scholarly topics.

made the greatest contribution to a specific topic. As an example, each year the U.S. National Library of Medicine database of biomedical literature (MEDLINE) is growing by approximately 700,000 articles from 21,000 different journals and as of 2015 contains reference data on over 25 million resources.[1] While this abundance of resources provides an incredible opportunity, it can also overwhelm researchers, whether they are searching for information as part of learning a specialization or trying to develop a broad understanding of different fields.

While the current data deluge has exacerbated the problem of retrieving relevant documents, the problem itself is not new. The field of information retrieval has been proposing mostly topical solutions to the problem of retrieving relevant documents since the 1950s. At the same time, the field of bibliometrics/informetrics/scientometrics started harvesting "vast quantities of knowledge about knowledge, or metaknowledge" (Evans & Foster, 2011; p. 721) from journal articles. While the two subfields of information science mostly developed in parallel (Glänzel, 2015; Wolfram, 2015), more recently there have been efforts to bring these two subfields closer in ways that would benefit both (e.g., Glänzel, 2015; Mayr & Scharnhorst, 2015; Mutschke & Mayr, 2015; White, 2007a, 2007b, 2015; Wolfram, 2015). This paper aims to contribute to those efforts, not only by utilizing knowledge from both subfields, but by proposing solutions for identifying related research that would be useful in building information systems and delineating reference sets for informetrics research.

In this research, we propose a topic evolution tree (TET) that builds on prior research to present different evolutionary paths for topics to different individuals, based on the context that is relevant to their research. To achieve this we use a heterogeneous bibliographic network (Sun, Han, Yan, Yu, & Wu, 2011) constructed from four types of entities present in a scientific papers repository: papers (P), venues (V), authors (A), topics or keywords (K), and the relationships between them. In TET we utilize multiple meta-paths[2] between topic nodes in the heterogeneous graph to identify the topics that a given topic has evolved from. In constructing the TET for a topic, we calculate a score for each meta-path instance based on the edge weights of the relationships. This approach shares similarities with the calculation of path instance scores as proposed in PathSim (Sun et al., 2011). However, we propose that for topic evolution, better performance results can be obtained by adding meta-path restrictions based on a new "contribution" edge as proposed for citation recommendation (Liu, Yu, Guo, & Sun, 2014b). Namely, we use not only the edges previously used in bibliographic networks, such as "written by", "cited by", "published in", or "used (topic)" (Lee & Adorna, 2012; Shi, Kong, Yu, Xie, & Wu, 2012; Sun et al., 2011;), but also contribution edges derived from PageRank calculations for each type of node (Liu et al., 2014b). For example, the "contributed-by-author" edge is calculated for each topic over a graph of authors citing other authors. We employ contribution edges as a means to restrict the context of a meta-path so that a random walk of the heterogeneous graph generates a TET showing the evolution of a topic in a particular context. In other words, we obtain the evolution of each branch of the TET based on the context or query topic that is specifically of interest to each scholar. The workflow used to generate a topic evolution tree is shown in Fig. 1.

As an example, in the information retrieval domain, a researcher might be interested in the evolution of the topic "Cloud computing"[3] (user query) and the papers, authors, or venues that made the most significant contribution to the evolution of that topic. The topic "Cloud computing" would be the root node of the TET and each edge from that root to a child node represents the evolution of "Cloud computing" from a contributing topic. One of the topics contributing to "Cloud computing" is the Big Data topic "MapReduce", so there would be an edge in the TET from "Cloud computing" to a child node labeled "MapReduce". Since the restricted meta-path uses the contribution edge from the bibliographic network, the subtree in the TET for "MapReduce" focuses on the evolution of that topic in the context of "Cloud computing". Thus, the evolution of the topic "MapReduce" could be different in the context of cloud computing than in the context of data security, since the papers, authors, or venues covering "MapReduce" will have made at least slightly different contributions in one context versus the

[1] http://www.ncbi.nlm.nih.gov/pubmed.

[2] A meta-path P is a path defined between two nodes on the heterogeneous graph and specifies the types of nodes and relationships in the graph that are on that path (Sun et al., 2011).

[3] We select this topic as a case study because it is a relatively new and important topic in the ACM Digital Library and in a domain that we are familiar with. In the future, we plan a more comprehensive study to evaluate the quality of TETs for a larger number of randomly selected topics.

other. It should be noted that this does not imply that the history of a topic such as cloud computing or MapReduce cannot be ascertained, but that the contribution edge added to a heterogeneous graph can be used to generate a TET containing the history of a topic from the viewpoint of the context of interest to the user. In the above example, the evolution of MapReduce would be included only in the context of its contribution to cloud computing. Conceptually this is similar to the research by Small, Boyack, and Klavans (2014) where the emergence of topics can be viewed from different levels of granularity.

2. Related work

In this paper we argue for the usefulness of restricted meta-paths over heterogeneous bibliographic networks both for finding highly relevant information and for developing an understanding of how a scientific topic has evolved. Such bibliographic networks are one example of a heterogeneous information networks (HINs) that capture the semantics of a real-world network (Sun & Han, 2012). Research on mining HINs has explored clustering, classification, and relationship prediction, but the earliest work on using meta-paths over HINs was PathSim (Sun et al., 2011), which examined the use of meta-paths for similarity search. An example of a meta-path over a bibliographic network that links two authors who both publish papers in the same venue is:

$$\text{author} \rightarrow \text{paper} \rightarrow \text{venue} \leftarrow \text{paper} \leftarrow \text{author}$$

In PathSim, a bibliographic database is used to calculate the similarity between peers in terms of similar publishing profiles (number of papers published in the same venues). To calculate similarity, meta-paths were specified using a network schema that defines the types of nodes and edges that exist in the network. The calculation of similarity in a HIN has also been extended to heterogeneous objects (Shi et al., 2012) for purposes of calculating the relevance between objects. Instead of calculating whether two authors are peers in that they have similar publishing profiles, comparing dis-similar objects could be used to identify the publishing venues most relevant to an author. Meta-paths over an HIN have also been explored for link prediction (Sun, Han, Aggarwal, & Chawla, 2012) and recommender systems (Yu et al., 2014). Although these different applications varied in their use of symmetric or asymmetric meta-paths, restrictions on the meta-path were first applied by Liu, Yu, Guo, Sun, and Gao (2014a) for citation recommendation.

Although meta-paths are a recent development, the field of information retrieval has been addressing the issue of identifying the topicality of documents for the purposes of identifying relevant literature for a very long time. Controlled vocabularies have been used as a primary vehicle to "optimize the precision and recall capabilities of a retrieval language" (Svenonius, 2003, p. 823). Much of the early work in this area was document-centered, attempting to find best ways to identify the topic of a document. However, a number of researchers have argued that a much more useful approach to knowledge organization and representation would be one focused on users and their needs, i.e., be user-centered (e.g., Bates, 1998; Fidel 1994; Harter, 1992; Hjorland, 1992; Mai 2001; Wilson, 1968). Such an approach follows from the idea that the relevance is not necessarily defined by finding documents on a particular "subject", but about finding documents that meet the user's need in a broader sense (e.g., Harter, 1992). This paper is an attempt towards such user-centered indexing that starts from the subject of the document, but utilizes traces of its usage by a community of researchers in order to anticipate its usefulness to an individual user.

There has also been a significant body of research focused on topic extraction in both information retrieval and informetrics that leverages refinements to the Latent Dirichlet Allocation (LDA) model (Blei, Ng, & Jordan, 2003), using a bag of words approach to discover topics in a text corpus of documents (commonly the title and the abstract). As noted by Thuc and Srinivasan (2008), early approaches to topic extraction, such as cluster models (Liu & Croft, 2004), described documents by a single topic, but subsequent research has focused on identifying multiple topics, based on LDA or its extensions. For example, researchers have used citation relationships either to boost the weight of cited topics (Jo, Lagoze, & Giles, 2007) or as extensions to the LDA model (such as the inheritance topic model (He et al., 2009) and dynamic topic model (Blei & Lafferty, 2006)). A subset of this research has focused on topic evolution in general, and the emergence of new topics in particular. An example of a system designed to identify emerging topics is the BioJournalMonitor (Mörchen et al., 2008). This system uses a Topic-Concept model that extends LDA by mapping extracted topics to concepts in the NLM's Medical Subject Headings (MeSH) ontology in order to identify topics that could be added to the ontology. Newman, Karimi, and Cavedon (2009), on the other hand, applied LDA to the results of MeSH-based queries on PubMed, in order to help users interpret MeSH terms. In another example, Griffiths and Steyvers (2004) applied LDA to papers in the Proceedings of the National Academy of Sciences (PNAS) in order to identify "hot" topics as a means of targeting scientific funding. The use of LDA to assist science policy makers in identifying research areas in need of funding has also been recently examined from the perspective of stem cell research (Wu, Zhang, Hong, & Chen, 2014). LDA was extended in the author-topic model (Steyvers, Smyth, Rosen-Zvi, & Griffiths, 2004) to enable queries to identify topics by author or authors by topic. This was further extended (Bolelli, Ertekin, Zhou, & Giles, 2009) to include extracting topics by time segment, with each segment building on the prior ones. This approach improved the performance of identifying topic trends over time as emerging (hot topics) or declining (cold topics).

A number of studies have focused on combining LDA with citations. For example, Mann et al. (2006) proposed the Topical N-Grams (TNG) extension to LDA in order to extract topics based on phrases and examine the citation relationship between those topics in order to determine their longevity. The use of citations to extend LDA was also incorporated into

the generation of topics in the citation-aware inheritance topic model (ITM) by He et al. (2009). That approach extended LDA by separating the topic space into an inherited part (based on citations) and an autonomous part generating new topics, which conceptually shares similarities with the seminal work by Price (1965) which found that scientific research citations consist of "classic" publications and an ephemeral research front of recent publications within a scientific community. Later research on topic extraction has employed LDA not only for topic evolution, but also in other applications, such as citation prediction (Yan, Tang, Liu, Shan, & Li, 2011). The citation graph was used by Jo et al. (2007) to address a known drawback of LDA that topics can include non-topic terms (He et al., 2009).

The evolution of science and the identification of emerging scientific trends have also been studied from the perspective of decades of research on citation networks. Early work by Garfield, Sher, & Torpie, 1964 examined the citation network of key publications in the discovery of the DNA code. This approach, known as algorithmic histography, is used in HistCite (Garfield, Pudovkin, & Istomin, 2003) to generate an account of the topics composing a scientific specialty. Based on bibliographic data for a collection of works in the domain of interest, HistCite identifies other papers as outer references that are potentially important to the evolution of a topic. Additional bibliographic data can then be loaded in an iterative process based on the outer references. For an example of the HistCite approach as applied to bibliographic coupling and co-citation, see Garfield (2001).

Citation networks are also leveraged in CiteSpace (Chen, 2006), which builds on Price's concept of a research front (Price, 1965). The citation network is used to identify pivotal points in the evolution of a research topic. Similar to the LDA extensions by He et al. (2009) in the inheritance topic model, and the domain-specific BioJournalMonitor (Mörchen et al., 2008), a primary goal of CiteSpace is the identification of emerging topics. Co-citations in the network are used to identify clusters that are labeled based on word-terms. The labeling of co-citation clusters based on research topics shares similarities with the inclusion of keyword nodes in the TET, and the use of a heterogeneous network of terms and articles in CiteSpace is identified by Chen as providing a comprehensive representation of the dynamics of a specialty. In contrast to the building of TETs, CiteSpace identifies clusters in the citation network and builds word terms based on those clusters. Similar to HistCite, processing in CiteSpace starts with a download of bibliographic data from the Web of Science (WoS) or PubMed based on a broad term search.

The algorithmic histography approach described by Garfield et al. (2003) has more recently been applied in CitNetExplorer (Van Eck & Waltman, 2014a) which aids researchers in understanding how a research area developed over time by visualizing a direct citation network. In contrast to HistCite, CitNetExplorer can handle much larger networks containing millions of publications, but still be understandable through the transitive reduction of the citation network and the ability to iteratively drill down and/or expand the network. Similar to CiteSpace, CitNetExplorer has the capability to identify clusters, but it uses the more recent smart local moving (SLM) algorithm (Waltman & Van Eck, 2013) to identify research community clusters. In contrast to both CiteSpace and the TET we present here, CitNetExplorer uses a homogeneous direct-citation network instead of a heterogeneous network including topics, so the focus of the visualization is on publications and not research topics. Although research communities within the citation network are identified using SLM and color-coded in the resulting visualizations, the communities are not labeled.

Recent research by Small et al. (2014) has combined direct-citation networks and co-citation to calculate an "Emergence Potential" (EP) for topics based on clustering a direct citation network to identify clusters that had articles that are new to the co-citation model and did not have articles in prior years. In contrast to much of the earlier work on direct citation networks which is retrospective, their research shares similarities with the previously mentioned LDA-based research in that its goal is to identify newly emerging topics.

Many of the previously proposed models start at a point in time with an identified topic or core corpus of documents and then identify new topics that emerge in subsequent temporal slices, by examining whether the emerging topics share similarities with the existing ones (e.g., Bolelli et al., 2009). Furthermore, many existing topic models are unsupervised algorithms, e.g., LDA and LSI approaches, making it difficult to interpret topic labels and define the number of topics, thus making the analysis and interpretation of the results difficult (Jiang, Liu, & Gao, 2015; Liu, Zhang, & Guo, 2013). In addition, most existing models do not take the context of the user into account, i.e., they do not account for the fact that seemingly identical topics might have had different development paths in different fields, subfields, or even schools of thought. The topic evolution tree, which we propose in this paper, fills that gap.

3. Data and methods

In this study, we investigate the topic evolution tree (TET) generation problem. Given a topic k^* of interest to a researcher and one of the K topics covered by papers in a scientific papers repository (in this case, a user query), a number of topics covered in the repository can contribute to k^*. Each of those topics can in turn be contributed to by other topics. Based on this kind of relationship, we can generate a tree structure for a query on the evolution of topic k^*, where k^* is used as the root of the tree. We define this structure as a TET (an example is depicted in Fig. 2). The most important step in generating a TET is defining the function for calculating the relationship between two topics K_i and K_j, i.e., $R(K_i, K_j)$. Two alternatives for this function are: (1) a simplified scenario in which we ignore the context of the user's original query k^* below the first level of the tree, and (2) a more sophisticated approach in which the user's original query is taken into account as the context for

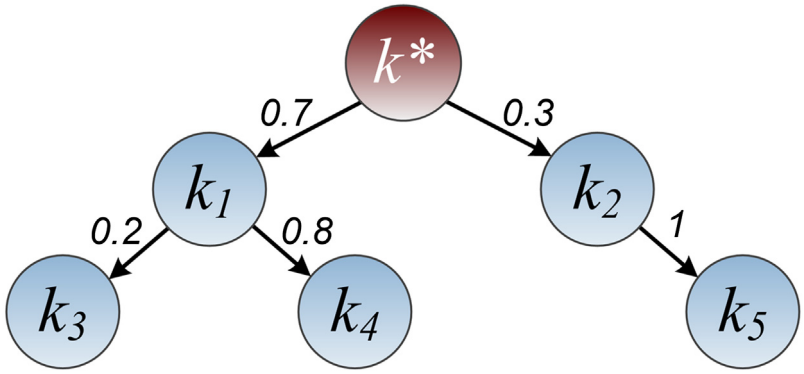

Fig. 2. Topic evolution tree for topic k* (depth = 2, maximum number of children = 2).

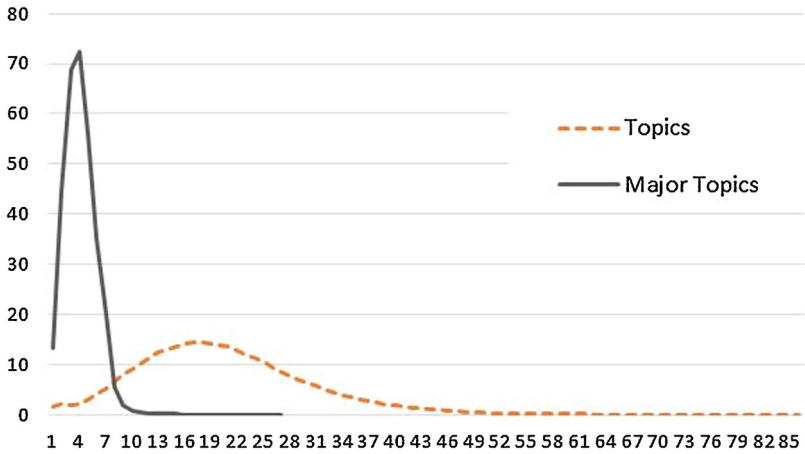

Fig. 3. Distribution of the number of topics and major topics per paper across the PubMed Central corpus.

the conditional relationship between subsequent topics, R(K_i, K_j| k^*). We apply both approaches to bibliographic networks from the biomedical science and computer science domains. The second approach is used to construct a query-centric TET.

3.1. Data

There are two data sources used in this study: the National Center for Biotechnology Information's PubMed Central (PMC) repository and The Association for Computing Machinery (ACM) Digital Library.

In 2014 we downloaded a snapshot of the metadata and full text for the collection of biomedical publications in PMC. The dataset contains 518,026 papers; each as an XML document. The data downloaded conforms to the National Library of Medicine's (NLM) Journal Archiving and Interchange DTD (Beck & Randall, 2013). Each paper is curated with multiple terms from the NLM's MeSH controlled vocabulary (Nelson, 2009) consisting of over 27,419 keyword descriptors (as of 2014), which can be further constrained by 83 qualifiers. The hierarchical nature of MeSH results in unqualified terms near the top of the hierarchy that are often too broad to provide guidance as to the topic of a paper (Cheung, Ouellette, & Wasserman, 2012). The metadata for each paper often contains multiple MeSH descriptors as topics. As shown in Fig. 3, the median number of topics for each paper is 19, but the median number of major topics is only 4. The major topics are generally more focused and semantically meaningful (lower in the hierarchy or qualified), so only the major topics are considered in this analysis. Since the qualified descriptors used as major topics are more narrowly focused, there are over 120,000 unique topics in the PMC dataset. The distribution of topics has a very long tail with many used in only a few articles. For this study, we analyzed the 3000 qualified descriptors most frequently used as major topics, so each qualified descriptor included as a topic in a TET is used to describe at least 50 papers in PMC. The MeSH vocabulary was also downloaded from PubMed to provide labels for each descriptor and qualifier, as well as the year that each descriptor was added to the MeSH vocabulary. This enabled the comparison of TETs for recent as well as established terms in the vocabulary. The publication venue for each paper is identified in each paper's metadata and is based on a controlled vocabulary of journal IDs. There are 3090 unique publication venues in the dataset.

The data were first imported into a relational database to generate import files for the graph database. As can be seen in Table 1, the number of nodes for research papers in the graph is slightly lower than the number of raw files initially

Table 1
Vertices and edges in the PMC and ACM heterogeneous graphs.

Vertices and edges	PMC	ACM
Vertices		
Paper	470,638	248,893
Author	155,050	387,932
Venue	3090	709
Topic	3,000	910
Total number of vertices	631,778	641.444
Metadata relationship edges		
"Written By" (paper to author)	720,790	632,074
"Published At" (paper to venue)	470,638	247,030
"Relevant" (paper to topic)	596,181	804,074
Citation (between papers)	693,150	755,162
Co-author (between authors)	2,156,604	1,379,240
Derived contribution edges from topic to:		
Paper	1,460,092	660,037
Author	299,100	449,417
Venue	149,450	224,863
Total number of edges	6,546,005	5,151,897

Table 2
Direct relationships in a heterogeneous bibliographic network.

Relation	Description
$P \xrightarrow{w} A$	Paper written by an author
$P \xrightarrow{p} V$	Paper published in venue
$A \xrightarrow{co} A$	Co-author relationship
$P \xrightarrow{c} P$	Paper citing a paper
$P \xrightarrow{r} K$	Paper relevant to keyword/(topic)

downloaded from PubMed Central. This is because of (1) a small number of duplicates, and (2) a small subset of articles (less than 10%) that do not have a PubMed ID (PMID). According to the NCBI Handbook (Maloney, Sequeira, Kelly, Orris, & Beck, 2013), records without a PMID represent book reviews and a few types of material not included in PubMed. These documents neither had references nor tended to be cited, so they were omitted. Without citation edges, such publications would not be traversed in walking the graph via the evolutionary meta-paths.

Initial downloading and processing of the ACM data was similar to the PMC data in that it was made available in an XML format that was parsed and loaded into a MySQL database for initial cleaning and preparation for generating the graph database import files. The ACM dataset contains 248,893 papers covering the period from 1960 to 2011, and most of the publications in the dataset contained metadata regarding both the venue and keyword information. Unlike the PMC data in which the major topics of each paper are identified based on qualified descriptors from the MeSH vocabulary, topics for publications in the ACM data are largely author-provided (although some computer science publications use a controlled vocabulary), which results in somewhat noisier keywords for ACM data. The distribution for the number of topics assigned to each paper in the ACM data is very similar to major topics in the PMC data; there is a sharp spike followed by a sharp drop in topics per paper. The ACM data contains a median of five topics per paper.

Similar keywords in the ACM data were merged using tokenization and Porter stemming algorithm; e.g., "K-Means", "K Means", "k means", and "K-Mean" were combined into a single topic. Additionally, keywords that appeared fewer than 10 times in the dataset were filtered out. After this initial data cleaning, there are 3910 unique keywords. As shown in Table 1 there were 709 unique venues and 387,932 distinct authors in the ACM dataset.

3.2. Unrestricted meta-paths TET

The use of meta-paths on a heterogeneous bibliographic network is a relatively new area of research (Sun & Han, 2012) compared to the mining of homogeneous graphs such as those based only on the citation relationships between papers. The merits of using a heterogeneous graph are twofold. First, a heterogeneous graph captures multiple kinds of information, e.g., authorship, venue information, and topic information. Second, by using different kinds of navigation methods, i.e., meta-paths, scholars can characterize different types of relations between target topics, which can be potentially important for scientific knowledge discovery. Although heterogeneous graphs provide an advantage over homogeneous graphs, most research on meta-paths over heterogeneous graphs has included only the direct relationships extracted from the source bibliographic data. We first construct a TET using an unrestricted meta-path over a heterogeneous graph for a bibliographic network based on the direct relationships presented in Table 2.

A heterogeneous graph based on direct relationships is shown in Fig. 4 and can be expressed as $T_G = (\mathcal{N}, \mathcal{R})$, where $\mathcal{N}$ is the type of nodes (Paper, Keyword, Author, and Venue) and $\mathcal{R}$ the possible relationships between nodes. Table 2 lists the

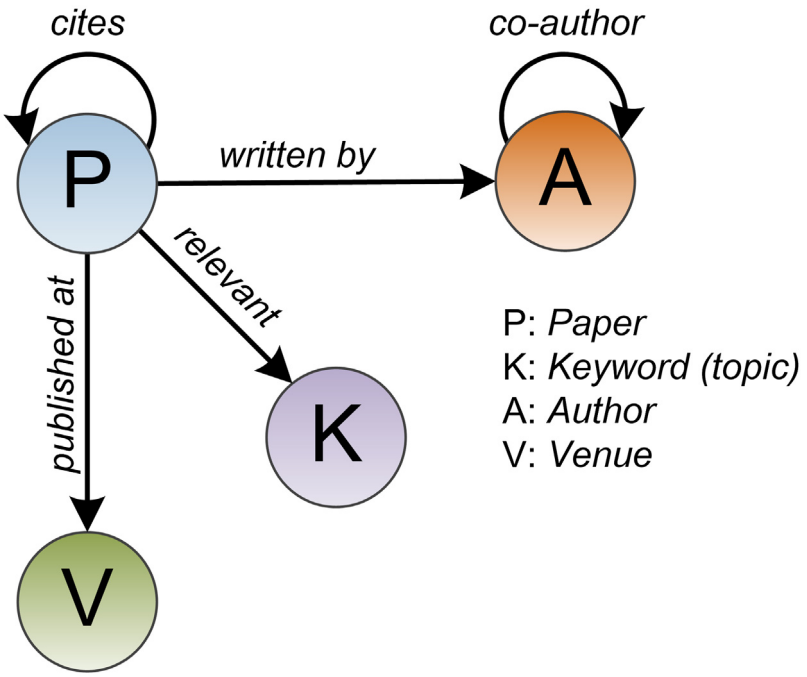

Fig. 4. Heterogeneous graph based on direct relationships.

direct relationships, $\mathcal{R}$, which include those between a paper and its authors (written by), its venue (published in), its topics (relevant to), and other papers (cites), as well as the relationship between authors (co-author). Each edge in the graph is assigned a non-zero weight, which represents the strength of the relationship. For each vertex in the heterogeneous graph, the edge weights for the same type of outgoing edge are normalized. For instance, the weight of the *relevant* edge from $paper_i$ to $keyword_j$ is defined as $w\left(p_i \xrightarrow{r} k_j\right) = 1/d\left(p_i \xrightarrow{r} K\right)$ where $d\left(p_i \xrightarrow{r} K\right)$ is the total number of keywords used to describe paper p_i. For the PubMed data we utilized the 3000 qualified descriptors that are most frequently used as primary topics, so K is restricted to those 3000 topics. The weight of the *written by* edge $P \xrightarrow{w} A$ between a paper and an author, the *co-author* edge $A \xrightarrow{co} A$ between two authors, and the *citation* edge $P \xrightarrow{c} P$ between two papers are all defined similarly. Since a paper can only be published in one venue, the weight of the $P \xrightarrow{p} V$ edge, representing a paper's relationship with the venue in which it is published, is always 1. Such heterogeneous graphs are constructed for both data sources (PMC and ACM).

When generating a TET, the root node k^* is the keyword node from the heterogeneous graph that identifies the topic of interest to a user from an evolutionary viewpoint. Walking down the tree; the inclusion of k_1 and k_2 as children of k^* indicates the extent of their contribution to the evolution of k^*. In the TET; the inclusion of k_3 and k_4 as children of k_1 shows those topics contributed to the evolution of k_1; given that k^* evolved from k_1.[4] The probability of an evolutionary relationship between two topics; as represented by each branch in the tree; is based on the conditional random walk of a meta-path on the heterogeneous graph; so $P(k_i \rightsquigarrow k_j|k^*) = RW_P(k_i \rightsquigarrow k_j|k^*)$. By definition; a meta-path does not need to start and end with the same type of node; and in some applications such as entity recommendation (Yu et al., 2014) they do not. However; in other applications such as clustering (Sun, Norick et al., 2012) and path similarity (Sun et al., 2011); a meta-path starts and ends with the same type of node. In applying meta-paths to topic evolution; each meta-path starts and ends at a keyword node.

The unrestricted meta-path based on citation relationships, which we use as a baseline for topic evolution, is $k^? \xleftarrow{r} P \xleftarrow{c} P \xrightarrow{r} k_i$, which represents the evolutionary path from a keyword k_i to papers described by that keyword (the paper has been identified by the author or PMC curators as *relevant* to the topic), to earlier papers cited by that paper, and then to the keywords $k^?$ used to describe such earlier papers. When starting at the root topic, the keyword k_i in the meta-path is k^*. The weight for each possible path over the heterogeneous graph between two topics is the product of the relevance and citation weights of the relationships in the meta-path. The strength of the evolutionary relationship between two keywords is the sum of the weights for all of the paths found between those two keywords through different citation relationships.

[4] The children in the tree are the topics a topic evolved from. This may sound counter-intuitive since they are child nodes in the tree, but they are not child topics in the evolutionary sense.

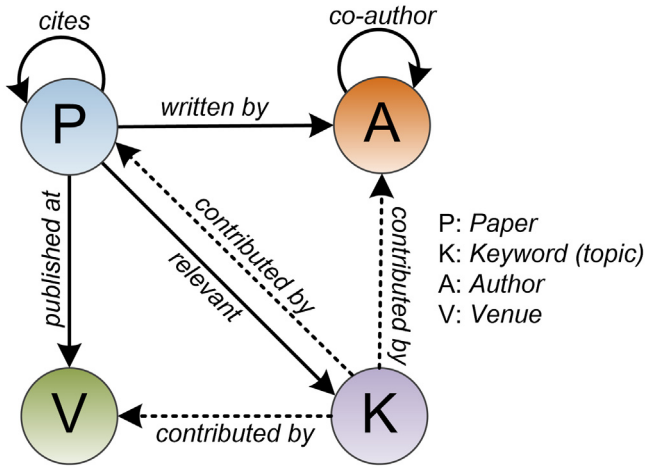

Fig. 5. Heterogeneous graph including the derived contribution relationship between topics and papers, authors, and venues.

Table 3
Derived contribution relationships in the heterogeneous graph.

Relation	Description
$K \xrightarrow{con} P$	Keyword/(topic) contributed to by paper
$K \xrightarrow{con} A$	Keyword/(topic) contributed to by author
$K \xrightarrow{con} V$	Keyword/(topic) contributed to by venue

Similar to the baseline based on the citation relationship, meta-paths on the heterogeneous graph of Fig. 4 could also be based on the authorship and co-author relationships, such as $k? \xleftarrow{r} P \xrightarrow{w} A \xrightarrow{co} A \xleftarrow{w} P \xrightarrow{r} k_i$ which is a meta-path through papers written by co-authors, or $k? \xleftarrow{r} P \xrightarrow{p} V \xleftarrow{p} P \xrightarrow{r} k_i$ which is the meta-path representing papers published in the same venue.

Since research builds on prior research, most topics will have at least a minimal probability of evolving from many other topics through the citation edges in the TET using the baseline meta-path $k? \xleftarrow{r} P \xleftarrow{c} P \xrightarrow{r} k_i$. This results in each node of the TET having many keywords as potential child nodes in the tree (topics they evolved from), and the same topic could also contribute to the evolution of multiple topics in the tree. Since the purpose of the TET is to present the evolution of the selected (root) topic, two restrictions are imposed: (1) once a topic is included in the TET, it will not be included again as contributing further down in the tree, and (2) the TET tree is pruned to impose a maximum on the number of child branches from each node. The first restriction ensures that the TET is represented as a tree and not a graph, which is visually easier for a user to comprehend. Also, since the TET is displaying the evolution of the specific root topic (and not the evolution of science in general), if a topic is higher in the tree, it made a greater contribution directly to the root topic and its subsequent contribution to another earlier topic is of less importance in the evolution of the root topic and is omitted. The second restriction limits each level of the TET to only showing those topics that made the greatest contribution to the topic of interest to the user. In the example of Fig. 2, the number of children is limited to 2.

3.3. Restricted meta-paths and query-centric TET

Since the unrestricted meta-paths TET uses meta-paths based only on direct relationships, it suffers from the limitation that the inclusion of topics in the TET below the first level is not strongly influenced by or relevant to the root topic of interest to the user's initial query. Each branch in the tree mainly reflects the direct evolutionary relationship between two levels in the tree without considering the context of interest to the user. As an example, the ACM data includes the topics "parallel processing" and "e-Science" which are both influenced by the topic "MapReduce"; using the baseline meta-path to build their respective TETs results in the same sub-tree under MapReduce in both TETs. However, the topics which contribute to the evolution of MapReduce vary when viewed from the context of the evolution of "parallel processing" or "e-Science". To address this limitation of the TET, three additional types of edges are added to the heterogeneous graph to characterize the contributions of papers, authors, and venues to each topic as shown in Fig. 5. In contrast to the direct relationships in the heterogeneous graph of Fig. 2 based on the metadata for each publication, the contribution relationships listed in Table 3 are derived edges that are not captured directly in the bibliographic data, but instead are calculated for each topic using the PageRank with Priors algorithm (White & Smyth, 2003). The inclusion of contribution edges in a heterogeneous graph was first applied to citation recommendation (Liu et al., 2014b) and is extended in this paper to topic evolution.

PageRank is often described as a Markov chain representing a random surfer on the Web who transitions in a stochastic manner to the next node in the Web graph based on properties of the current node. The importance of each webpage is

then based on the amount of time spent at that page's node in the graph. As noted by White and Smyth (2003), although the Markov chain analogy may be somewhat strained in the case of graphs other than the Web, PageRank is widely accepted as providing a useful definition of importance across other graphs such as bibliographic networks. As defined by White & Smyth, PageRank with Priors extends PageRank by defining a vector of prior probabilities $P_R = \{p_1, \ldots, p_{|V|}\}$ denoting the relative importance of each node, where some nodes are assigned a prior bias and others have a prior of 0. At each step of the calculation PageRank with Priors transitions back to one of the nodes with a prior bias based on a back probability β, $0 \leq \beta \leq 1$.

The contribution of each paper, author, and venue in the heterogeneous graph is calculated using PageRank with Priors by starting with classical homogeneous graphs where each vertex is of the same type (paper, author, or venue), and the citation relationship between vertices is utilized to calculate the PageRank score. The priors for each vertex are determined by first calculating a topic prior vector for each node. For example, in the paper graph each paper has a topical prior distribution $P(k_i|paper)$ which is the probability that a paper is relevant to topic k_i (the topic is labeled by keywordk_i). If a paper in PubMed Central used four of the 3000 qualified MeSH descriptors as primary topics, each of those four topics would have a prior probability of 0.25 for that paper. For the author graph, the topical prior distribution is $\sum P(k_j|paper_i)$, where $paper_i$ is published by the target author. Since the prior for an author is based on the sum of the probabilities for each qualified descriptor used as a primary topic across all of the papers written by that author, those topics on which researchers publish more frequently will have a higher prior. The PageRank with Priors calculation is run for each topic, for each contribution type (contribution of a paper, author, or venue to a topic). The result of the PageRank calculation is the topic authority vector for the topic. For example, the PageRank calculation for the contribution of papers to topic k_j generates the paper authority vector$Authority(paper_i|k_j)$ which contains the authority score for each $paper_i$ for a given topick_j.

Note that $topic_j$ being contributed to by $paper_i$ ($K_j \xrightarrow{con} P_i$) does not necessarily mean $paper_i$ is relevant to $topic_j$ ($P_i \xrightarrow{r} K_j$). The relevance of a paper to a topic results from the direct relationship of an author or curator identifying it as a major topic of the paper. In contrast, a paper could contribute to a topic even if the author did not specifically identify it as such. For example, some "*HLA-DRB1*" (a type of gene) papers can be important (e.g., contribute) to the topic "*Type* 1 *diabetes*" even if "*Type* 1 *diabetes*" is *not* the topic of the paper. The contribution relationship can be forward-looking in that a paper can make a significant contribution to a topic that had not yet emerged at the time the paper was written.

By including this novel contribution edge, the TET can capture the query-centric relationships between the target topic and the topics at every branch of the TET. In this approach, the contribution edges place restrictions on both the citing and cited papers in the baseline meta-path $k^? \xleftarrow{r} P \xleftarrow{c} P \xrightarrow{r} k_i$ to adjust the weight of the meta-path between two topic keywords based on the contribution to the target topic of the TET by: (1) both papers, (2) the authors of both papers, or (3) the shared publishing venue of the papers.

For the example TET shown in Fig. 2, for each paper described by the topic k_1 that cites a paper described by topic k_3, instead of the result being based only on the weight of the citation relationship and the relevance of the papers to their respective topics, the following restricted meta-path includes the contribution of each paper to the target topic k^* of interest to the user (the user's initial information need):

$$k^? \xleftarrow{r} \underset{con}{P} \xleftarrow{c} \underset{con}{P} \xrightarrow{r} k_i$$
$$\nwarrow k^* \quad \nwarrow k^*$$

Restricted meta-path based on the paper's contribution

For the TET shown in Fig. 2, if "MapReduce" is the topic represented by the keyword k^1, the topics included in the subtree under k^1 (e.g., topics k^3 and k^4) reflect not only the evolution of "MapReduce", but those topics in the evolution of "MapReduce" that also made the greatest contribution to the evolution of k^*.

The second type of restricted meta-path we propose is based on the assumption that if the author(s) of the citing and cited papers made significant contributions to the query topic, the generated TET will be more relevant to that topic. In this meta-path, the contribution of the authors (as restrictions) from both papers in the citation relationship to the query topic k* is used to calculate the weight for each path betweentwo topics:

$$k^? \xleftarrow{r} \underset{w}{P} \xleftarrow{c} \underset{w}{P} \xrightarrow{r} k_i$$
$$\searrow A \quad \searrow A$$
$$\underset{con}{\nwarrow k^*} \quad \underset{con}{\nwarrow k^*}$$

Restricted meta-path based on the author's contribution

Although each pair of papers in a citation relationship can have many potential paths based on this meta-path (since each paper may have many authors), the written-by relationship between a paper and its authors is normalized as discussed above and a paper with a larger share of the authors contributing to the target topic would result in a stronger evolutionary relationship between two keywords. The restricted meta-path based on the author contribution is less restrictive than the meta-path based on the paper contribution since over time authors are likely to publish on a greater number of topics than a single paper would identify as major topics.

Finally, we propose a restricted meta-path based on the contribution of the venue(s) to the query topic. The contribution to the target topic k^* by both of the venues in which the papers in the meta-path's citation relationship were published can also be used as a restriction on the baseline meta-path as follows:

Table 4
PubMed Central and ACM topics for which TETs were generated.

Topic	First used	Papers
PubMed central topics		
DNA-directed DNA polymerase—metabolism	2014	125
Epithelial cells	2014	76
Real-time polymerase chain reaction—methods	2014	79
Coral reefs	2010	78
Induced pluripotent stem cells—cytology	2009	111
G-Quadruplexes	2007	132
Chromatin immunoprecipitation—methods	2004	87
DNA replication	1999	426
Telomerase—genetics	1995	81
Receptors, nntigen, T-cell, alpha-beta—genetics	1991	82
Receptors, antigen, T-cell, alpha-beta—immunology	1991	86
Telomere—genetics	1991	107
T-lymphocyte subsets—immunology	1990	352
Polymerase chain reaction—methods	1990	767
CD4-Positive T-lymphocytes—immunology	1988	704
Flow cytometry—methods	1981	137
ACM topics		
Twitter	2009	129
Cloud computing	2008	357
Crowdsourcing	2008	115
MapReduce	2007	141

$$k^{?} \overset{r}{\leftarrow} P \overset{c}{\leftarrow} P \overset{r}{\rightarrow} k_i$$
$$\underset{\searrow}{p} V \underset{\searrow}{p} V$$
$$\underset{\nwarrow}{con} k^{*} \underset{\nwarrow}{con} k^{*}$$

Restricted meta-path based on a venue's contribution

Of these three restricted meta-paths, this is the least restrictive in that many journals and conferences publish research on a wide variety of topics in a domain, or across domains.

For selected topics in both the PMC and ACM datasets, the baseline meta-path and all three of the above restricted meta-paths based on the paper, author, and venue contribution to a topic were calculated for a selected subset of keywords.

4. Results

In order to validate the method proposed in this study, we constructed separate heterogeneous graphs for biomedical research (PMC) and computer science (ACM). The number of vertices and edges in each graph are listed in Table 1 as discussed in Section 3.1. The direct relationships, such as the authors of each paper, were extracted from the bibliographic metadata and the PageRank calculations used to generate the contribution edges were performed using version 2.0.1 of the JUNG graph library. The PageRank with Priors calculations were run for each of the three types of contribution edges for each topic. Each calculation was run for 500 iterations or until it reached convergence. The heterogeneous graphs were constructed using version 2.2.2 of the Neo4j[5] graph database.

We selected a set of 16 biomedical topics from the 3000 MeSH topics in the heterogeneous graph for the PMC data. Topics were selected to cover a range of different time spans since their first usage in the MeSH vocabulary (1981–2014). In the ACM data we identified the four most recently evolved popular topics, which have been used as keywords for over 100 papers in the corpus ("Twitter", "Cloud computing", "Crowdsourcing", and "MapReduce"). Each of these topics was first used as a keyword in 2007 or later. The selected PMC and ACM topics are listed in Table 4, along with the year each was first used and the number of papers in which it was used. TETs were generated and visualized for the baseline meta-path without restriction and the three restricted meta-paths based on the contributions of papers, authors, and venues to the target topic (the initial user information need). The product of the edge weights was calculated for each path in the graph that matched the meta-path being calculated and those products were summed by the endpoint topics of the meta-path. The TET was pruned to limit the number of outgoing evolutionary branches from each node to the five highest scoring branches. The TET was also pruned based on the score calculated for each branch in the tree. A low threshold was used since the score for a branch depends on the product of the relationships in the meta-path. The number of topics excluded at each node in the TET was logged. The resulting TET was written out as a JSON file and visualized in a web interface using Cytoscape.js (Shannon et al., 2003). The web-based visualization of the TET based on the paper contribution meta-path for the ACM topic "Cloud computing" is shown in Fig. 6. As the user clicks on a topic in the interface, the paper citation relationships that contributed

[5] http://neo4j.com/.

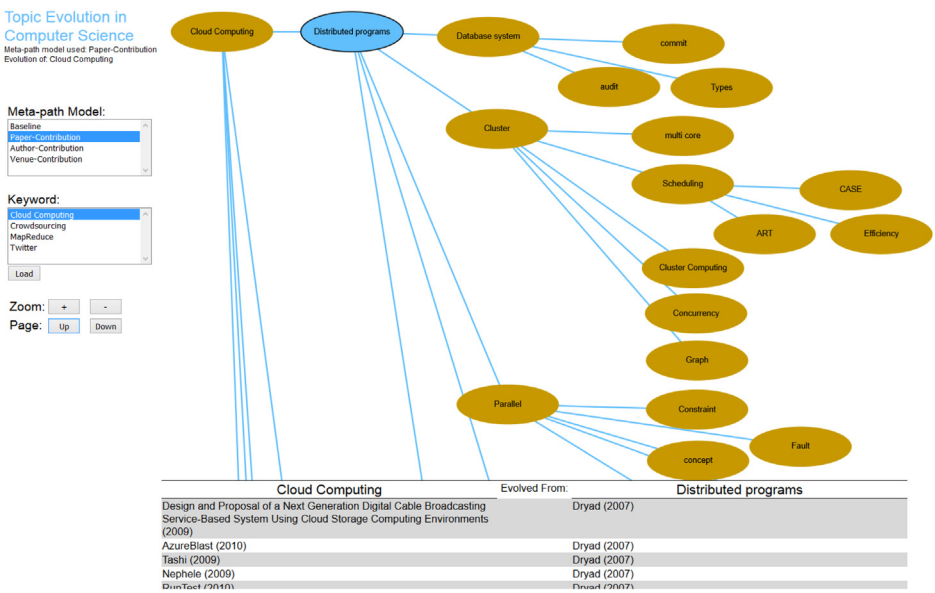

Fig. 6. Web-based visualization of a topic evolution tree (TET) for the ACM topic "Cloud computing" based on the paper contribution meta-path. The "distributed programs" topic is selected, so the citations displayed are those contributing the most to that topic being included in the evolution of the root topic.

most significantly to that evolutionary branch in the tree are listed. The visualizations for each of the selected PMC[6] and ACM[7] topics based on the baseline and all three restricted meta-paths are available on the web.

Although the citation relationship edges in the heterogeneous network are central to the meta-paths used to generate a TET, as shown in Fig. 6, the nodes in a TET represent topics and not publications. The graph nodes in most approaches to visualizing bibliometric networks represent publications, although CiteSpace (Chen, 2006) labels publication clusters based on the top keyword terms describing each cluster. For a summary of current approaches to visualizing bibliographic networks, see Van Eck and Waltman (2014b). A recent web-based interface that is closer to the visualization of a TET is Ariadne's Thread (Koopman, Wang, Scharnhorst, & Englebienne, 2015) which displays topics, authors, journals, and Dewey decimal classes related to a user's query. The goal of Ariadne's Thread differs from the TET in that instead of focusing on the evolution of a specific term, clicking on a node in the visualized graph generates a new search. As Koopman et al. clearly state, serendipity and the shifting of a user's interest as they explore the graph is a key characteristic of the Ariadne system.

There are no existing benchmarks for topic evolution (He et al., 2009) and different meta-paths can have different semantic meaning (Sun et al., 2011). For example, while the meta-path restrictions based on the contribution of papers and authors both focus on two papers in a citation relationship, the restrictions differ in how they moderate the importance of that citation relationship in the evolution between the topics at the endpoints on that meta-path. The paper restriction considers the extent of each paper's contribution to the topic of interest (root topic), whereas the author restriction is based on the contribution over time to the root topic by all of the authors for both papers. For topics below the root level in a TET based on a restricted meta-path, the evolution of each topic is also based on the context of the particular TET. For example, if the topic "File systems" were included in the TETs for two different topics, the topics included in the sub-trees under "File systems" in each TET would differ since their inclusion is moderated by their contribution to the root topic of their respective TET.

The TET for a specific topic is a snapshot of that topic's evolution to-date and will slowly evolve over time. In this sense, the evolution of a topic in a TET differs from the emergence of a topic. Although there may not be a strict definition of emergence, newness and growth are common themes (Small et al., 2014). For example, Small et al. (2014) identify cloud computing as emerging in 2010 and being one of the top 25 topics to emerge in the period 2007–2010. From a TET perspective, evolution is viewed as an ongoing process, so cloud computing will continue to evolve for years after its emergence. A longer term example is artificial intelligence (AI) which was a hot research topic in computer science in the 1970's but subsequently entered a period referred to as the "AI Winter" (Hendler, 2008) in which research decreased substantially. Due to recent developments in Deep Learning, AI is a hot topic again, but a TET on AI generated in 2015 would differ from a TET generated only on research prior to 1990.

Let us examine TETs generated using the baseline (i.e., unrestricted) meta-path and the restricted meta-paths based on paper and author contributions for the ACM topic "Cloud computing" (i.e., user query "Cloud computing" (Table 5)). The

[6] http://www.topicevolution.org/biomedical/TopicEvolution.html.
[7] http://www.topicevolution.org/computerscience/TopicEvolution.html.

Table 5
Comparing a snippet of the TETs generated for the ACM topic "Cloud computing" based on the baseline, paper contribution, and author contribution meta-paths. The number of child topics excluded at each node due to pruning is included in parenthesis.

Baseline	Paper contribution	Author contribution
Cloud computing (1199)	Cloud computing (194)	Cloud computing (93)
Grid (1371)	Distributed programs (1)	Conversation
Peer-to-peer networks (817)	...	Interviews
File sharing (749)	MapReduce (15)	Read (218)
...	...	Concurrency (227)
Hyperlink (887)	Grid	...
...	Data center (28)	Traffic classification
Anonymity (1224)	Hardware (103)	Locking (83)
...	Management (122)	...
Small world (618)	Virtual machine (63)	Network measurement (65)
...	...	...
DHT (520)	XEN (78)	File systems (181)
...	Virtualization (86)	Overlay network (214)
Layout (1795)	...	...
Automation (2263)	File systems (53)	Cluster (158)
...	...	...
Channel assignment (443)	Virtual machine monitor (4)	Redundancy (120)
...	...	...
Standard cell (544)		RAID (92)
...		...
Constraint (2530)		Log (178)
...		...
Sketching (1691)		Virtual machine monitor (137)
...		Disk scheduler
Virtualization (2264)		Virtualization (198)
...		...
File systems (1281)		Scalability (405)
...		...
MapReduce (973)		Paging (43)
...		...
Virtual machine monitor (558)		Operating system (160)

most significant difference in the trees at this level is that including only the five topics making the greatest contribution has required pruning significantly more topics in the baseline TET than in the restricted ones. For the baseline and paper contribution TETs, four of the five topics identified as making the greatest influence on the evolution of "Cloud computing" are the same: "Grid", "File systems", "MapReduce", and "Virtual machine monitor". Both of those TETs include "Grid" at that first level, and a review of the top ten citation relationships in that evolutionary path identified the same paper, *A Break in the Clouds: Towards a Cloud Definition* (Vaquero, Rodero-Merino, Caceres, & Lindner, 2009), as having the greatest influence. The abstract for that paper talks directly to the relationship between cloud computing and grid computing, "*This paper pays much attention to the Grid paradigm, as it is often confused with Cloud technologies. We also describe the relationships and distinctions between the Grid and Cloud approaches.*"

Since the first level of the TET represents the topics making the greatest contribution to cloud computing (as of the download of the ACM database in 2011), it would be expected that these topics would also be considered closely related to cloud computing in the Ariadne system. Entering "Cloud computing" as a search term in Ariadne's web interface[8] generates a graph of those terms, authors, and journals most closely related to cloud computing based on the OCLC ArticleFirst database. In addition to using a different bibliographic database as its source (ArticleFirst vs. ACM), Ariadne also uses a different approach to searching the heterogeneous bibliographic data in the database. Instead of citation relationships, Ariadne uses the cosine similarity of term vectors describing each entity after first using dimensionality reduction on the vectors to make the calculation tractable (Koopman et al., 2015). Despite these differences, three of the five terms included in the first level of the TET were identified as those most closely related to cloud computing (distributed computing, grid computing, and virtual machines/virtualized). A fourth term (MapReduce) was identified as being closely related to cloud computing through author relationships. Some differences would be expected since the TET's focus is on the evolutionary contribution of topics to cloud computing and Ariadne's focus is on related subjects.

In the TET restricted based on the paper's contribution, below the first level of the tree the subtree for the topic "Grid" is still focused on topics related to "Cloud computing" and a review of the top papers contributing to the evolutionary path of these topics in the TET ("data center", "hardware", "management", "virtual machine", "XEN", and "virtualization") reveals that two papers closely related to "Cloud computing" had the greatest influence on these paths: *Emergence of the academic computing clouds*, and *Xen and the art of virtualization*. "Virtualization" being a key concept in cloud computing.

[8] http://thoth.pica.nl/relate.

In contrast, the baseline meta-path diverges from cloud computing to include topics such as "hyperlinks", "anonymity", "automation", "sketching", "thumbnail images", "video", and "user-interface storyboarding", with the papers making the greatest contribution to those evolutionary paths having minimal to no relationship to "Cloud computing". This divergence from a clear evolutionary path for "Cloud computing" can be traced to two significant limitations of the baseline TET. First is that the edge relationships in the tree are not restricted to papers with any relationship to "Cloud computing" whereas the restricted TET requires an indirect relationship through the PageRank calculations for the contribution edges from papers in the bibliographic network back to the root topic ("Cloud computing" in this case). This results in the baseline TET having evolutionary paths such as:

Cloudcomputing → Grid → Peer-to-peernetworks(P2P) → Hyperlink → Thumbnail

Here the evolution of the topic "Hyperlink" from "Thumbnail" is motivated the most by an article titled *Using thumbnails to search the Web*; a human factors study of whether users are able to find information on the web faster using plain text, thumbnail images, or a combination of both. Although it has been cited 70 times, it does not address the needs of a user learning about cloud computing. In the PageRank calculations run to determine the contribution of papers to each topic, this paper contributed to hyperlinking, mobile interfaces, web browsing, and thumbnails, but had no relation to cloud computing, or even to grid computing or P2P networks which are topics on the evolutionary path to "Cloud computing" in the baseline TET. For the path in the baseline TET between "Hyperlink" and "Peer-to-peer networks", the papers identified as most influential contributed to topics such as "hyperlinking", "PageRank", "P2P networks", and "web graphs", but not "Cloud computing", or even "grid computing" - the next node in the path towards "Cloud computing".

In contrast to the baseline TET, each branch of the paper contribution TET contributes at least indirectly to the topic of interest to the user since the PageRank score is included in the product used to weight each branch of the TET. As an example, one of the evolutionary paths in the paper contribution TET is:

Cloudcomputing → Filesystems → Scale → Latency → Architecture

Each topic on this path is motivated by articles focused on scalability, distributed systems, distributed computations, fault tolerance and other topics related to cloud computing, but none of the papers is directly identified by its authors as relating to "Cloud computing" except for those on the last edge, from "File systems" to "Cloud computing". Many of the papers motivating nodes on this path predate the idea of cloud computing.

The second limitation of the baseline TET is that the same keyword can have inconsistent semantics as illustrated by the following path in the baseline TET:

Cloudcomputing → Grid → Layout → Automation

As noted above, the main paper influencing the path from "Grid" to "Cloud computing" in both the baseline and paper contribution TET is closely related to cloud computing and is highly cited. However, in the baseline TET the path from "Layout" to "Grid" is no longer influenced by cloud computing, and since it lacks the context provided in the restricted TET, the topic "Grid" has inconsistent semantics. Instead of referring to grid computing, the evolution from "Layout" to "Grid" is dominated by papers covering the layout of computer circuits at the hardware level that are unrelated to cloud computing, and arguably more accurately characterized as being computer engineering than computer science. This issue with the inconsistency of the topic semantics is more pronounced in the ACM data where keywords are assigned by the author and the most common meaning of a term can change over the course of decades. In the PMC data this is less of an issue since the MeSH controlled vocabulary is used, and although in a limited set of cases the same descriptor appears in different branches of the hierarchy, such terms are less likely to be used with the same qualifiers as major topics. The restricted meta-paths mitigate this problem since the PageRank calculations used to calculate the contribution of each paper to the root topic would result in any branch towards unrelated topics such as hardware circuit layouts being pruned when generating a TET on cloud computing.

The baseline TET also suffers from including too many topics. This limits the value of the baseline meta-path as an aid in learning a topic. The number of nodes in the baseline TET is restricted mainly by pruning to five directly contributing topics. This results in the baseline TET containing 3713 topics for "Cloud computing", whereas the paper contribution TET contains only 257 topics for "Cloud computing".

Restricted TETs were also created for each of the selected topics based on the author's contribution. Paths in this TET are influenced by the research history of the authors in the citation relationship. A paper that was important to the first three topics that "Cloud computing" evolved from in the author contribution TET is a highly cited paper titled *A view of cloud computing* that was published a year before the ACM data was collected and authored by a team of established researchers; three of whom contributed to "Cloud computing" based on the PageRank calculations. Similar to the paper contribution TET, both "File systems" and "Virtual machine monitor" are included as topics at the first level of the TET, "MapReduce" is included as a topic further down in the TET, but "Grid" is not included at all as a topic. This highlights a key difference between the paper and author contribution TETs and illustrates the different research needs addressed by each. While grid computing shares similarities with cloud computing, where one aim is to provide access to a highly scalable shared pool of resources (Mell & Grance, 2011), at least initially they addressed different communities. Grid computing focuses on providing heterogeneous computing resources to scientists through a virtual organization (VO) that often encompasses multiple government-funded super-computing facilities. Cloud computing grew out of a need in the commercial sector to provide scalable, low-cost,

homogeneous resources from a single provider. A comparison of the authors included in the PageRank results for "Cloud computing" and "Grid" reveals that less than 12% of the authors identified as contributing to "Cloud computing" are identified as contributing to "Grid" (as of 2011 when the ACM data was extracted). The paper and author contribution restricted TETs present different views as to the evolution of a topic; the paper contribution TET identifies those topics that contribute to the research itself, whereas the author contribution TET considers the research history of the authors contributing to a topic.

Restricted TETs based on the venue contribution were also run for the same selected keywords, but share many of the limitations seen in the baseline TET. At the first level of the tree, the venue contribution TET for "Cloud computing" shares four of the five topics included in the paper contribution TET, but overall contains a total of 3461 topics. Similar to the baseline TET, it contains topics unrelated to "Cloud computing", such as "photography", "digital rights management", and "digital pens". In comparison to the contribution of a paper to a topic, or even an author's contribution over time to a set of topics, research journals and conferences cover a broad range of topics and pose only a minimal restriction when compared to the baseline meta-path.

5. Discussion and conclusions

Recent research has applied the mining of relationships in heterogeneous graphs such as bibliographic networks to purposes such as path similarity, identifying emerging research topics, and citation recommendation. In this paper we examined mining heterogeneous graphs from different scientific domains to map the evolution of topics as a TET. Prior research has focused on heterogeneous networks composed of objects and relationships extracted directly from bibliographic metadata, and while we find that meta-paths over such networks can provide a baseline for comparison, too many unrelated topics are included in the resulting TET. By extending bibliographic networks with derived relationships that capture the indirect contribution of papers, authors, and venues to a topic, a TET can be generated using restricted meta-paths that constrain the evolution from earlier topics to the context of the root node in the TET. This approach is applied to both biomedical research (PMC) and computer science research (ACM) and then visualized in a web interface showing the papers or authors that made the greatest contributions along each evolutionary branch of a topic's TET.

An initial review of the TETs generated for the biomedical and computer science terms indicates that a TET based on the paper contribution restricted meta-path presents a good summary of the topics that have contributed to the evolution of the root topic. Although the biomedical terms selected from the MeSH vocabulary covered a wide temporal range, both the MeSH and ACM terms analyzed are frequently used. An open issue is how prominent a topic must be for the generated TET to accurately represent its evolution. As noted by Small et al. (2014), a similar limitation potentially exists with the emergence of topics in that smaller topics might be missed.

In the TETs explored in this paper, the context for the restrictions on the meta-path has been the topic that is the root of the tree, but the context could be another topic or a scientific community. For example, MapReduce is a key technology used in Big Data and TETs could be generated for that topic in the context of different research communities—such as researchers focused on parallel computing or e-Science. The resulting TETs would then show the evolution of MapReduce from the perspective of those particular scientific communities. Scientific communities could also be identified using community detection algorithms such as the SLM algorithm (Waltman & Van Eck, 2013) which is used in CitNetExplorer. Contribution edges could be calculated for each paper or author's contribution to the identified communities.

In this paper the TETs were constructed from data covering roughly 240,000 (ACM) and 470,000 (PMC) papers. An open question that we intend to study in future research is whether better TETs could be generated by including more data or more information about the context of each citation within the text of the paper. In the Big Data community it is often argued that a large volume of data trumps better models (Halevy, Norvig, & Pereira, 2009) and the full PubMed repository contains metadata (but not full text) for over 25 million papers on biomedical research. Alternately, full text could be used to provide the context of each citation, possibly allowing each citation to be weighted towards individual keywords used to describe a paper. Using the full text for a paper, more sophisticate methods could also be used to infer the topic(s) of each paper. For instance, an LLDA (Labeled LDA) algorithm combined with author contributed keywords could be used to infer a topic probability distribution as was used in Liu et al. (2013).

Another issue to be explored is how to present or evaluate the quality of the TETs for a large number of possible topics. The exemplar TETs generated for the selected biomedical and computer science topics discussed in this paper can be viewed in the web interface shown in Section 4, but for a large number of potential topics, an open issue from a human–computer interaction (HCI) perspective is how to allow a user potentially unfamiliar with a domain to select which of thousands of possible topics they wish to explore. A related aspect we plan to investigate is how the proposed TET can enhance the user experience for scholarly retrieval. Prior experience in search system interface and novel algorithm evaluation (Sutcliffe, Ennis, & Hu, 2000; Tamine-Lechani, Boughanem, & Daoud, 2010) show that the cost of such studies can inevitably be high. One possible approach is to invite a number of academic scholars to use and evaluate the TET system. While the cost of human evaluation is high, although the first level of the TET is similar to the results from the Ariadne system (Koopman et al., 2015), future work is required to further evaluate the quality of the TET being generated for a broad selection of topics.

The methods proposed in this paper open up a promising research path to utilize heterogeneous graphs of scientific papers to enhance the information retrieval of scientific literature as well as to identify research communities focused on particular lines of research.

Authors' contributions

Conceived and designed the analysis, collected the data, contributed data or analysis tools, performed the analysis and wrote the paper: Scott Jensen.

Conceived and designed the analysis, collected the data, contributed data or analysis tools, performed the analysis and wrote the paper: Xiaozhong Liu.

Collected the data and contributed data or analysis tools: Yingying Yu.

Wrote the paper: Staša Milojevic.

Acknowledgements

The authors wish to thank the National Center for Biotechnology Information for making bibliographical metadata on biomedical literature available through PubMed and also wish to thank the Association for Computing Machinery for making bibliographic metadata on computer science literature in the ACM digital library available for this project.

References

Barile, S., Franco, G., Nota, G., & Saviano, M. (2012). Structure and dynamics of a T-shaped knowledge: from individuals to cooperating communities of practice. *Service Science*, *4*(2), 161–180.

Bates, M. J. (1998). Indexing and access for digital libraries and the Internet: human, database, and domain factors. *Journal of the American Society for Information Science*, *49*(13), 1185–1205.

Beck, J., & Randall, L. (2013). NLM DTD to NISO JATS Z39. 96-2012. In *The NCBI Handbook [Internet]* (2nd edition). Bethesda MD: National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/books/NBK169004/ Accessed 28.08.15

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd international conference on machine learning, ACM*, 113–120.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, *3*(March), 993–1022.

Bolelli, L., Ertekin, S., Zhou, D., & Giles, C. L. (2009). Finding topic trends in digital libraries. *Proceedings of the 9th ACM/IEEE-CS joint conference on digital libraries (JCDL '09), ACM*, 69–72.

Bornmann, Ll., & Mutz, R. (2015). Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222.

Chen, C. (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.

Cheung, W. A., Ouellette, B. F. F., & Wasserman, W. W. (2012). Quantitative biomedical annotation using medical subject heading overrepresentation profiles (meshops). *BMC Bioinformatics*, *13*(1), 249.

Donofrio, N., Spohrer, J., & Zadeh, H. (2010). Research-driven medical education and practice: a case for T-shaped professionals. In *IBM working document.*. Last downloaded 02.08.15 from: http://www.ceri.msu.edu/wp-content/uploads/2010/06/A-Case-for-T-Shaped-Professionals-20090907-Hossein.pdf

Evans, J. A., & Foster, J. G. (2011). Metaknowledge. *Science*, *331*(6018), 721–725.

Falk-Krzesinski, H. J., Börner, K., Contractor, N., Fiore, S. M., Hall, K. L., Keyton, J., et al. (2010). Advancing the science of team science. *Clinical and Translational Science*, *3*(5), 263–266.

Fidel, R. (1994). User-centered indexing. *Journal of the American Society for Information Science*, *45*(8), 572–576.

Garfield, E. (2001). From bibliographic coupling to co-citation analysis via algorithmic historio-bibliography. Presented at Drexel University, Philadelphia, PA. http://garfield.library.upenn.edu/papers/drexelbelvergriffith92001.pdf. Last retrieved 11.01.16.

Garfield, E., Pudovkin, A. I., & Istomin, V. S. (2003). Why do we need algorithmic historiography. *Journal of the American Society for Information Science and Technology*, *54*(5), 400–412.

Garfield, E., Sher, I. H., & Torpie, R. J. (1964). *The use of citation data in writing the history of science*. Philadelphia: Institute for Scientific Information.

Glänzel, W. (2015). Bibliometrics-aided retrieval: where information retrieval meets scientometrics. *Scientometrics*, *102*, 2215–2222.

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

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, *24*(2), 8–12.

Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science*, *43*(9), 602–615.

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, C. L. (2009). Detecting topic evolution in scientific literature: how can citations help? *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09), ACM*, 957–966.

Hendler, J. (2008). Avoiding Another AI Winter. *IEEE Intelligent Systems*, *23*(2), 2–4.

Hjorland, B. (1992). The concept of subject in information science. *Journal of Documentation*, *48*(2), 172–200.

Jiang, Z., Liu, X., & Gao, L. (2015). Chronological Citation Recommendation with Information-Need Shifting. *Proceedings of the 24th ACM international conference on information and knowledge management (CIKM '15), ACM*, 1291–1300.

Jo, Y., Lagoze, C., & Giles, C. L. (2007). Detecting research topics via the correlation between graphs and texts. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '07), ACM*, 370–379.

Jones, B. F. (2005). The burden of knowledge and the 'death of the renaissance man': is innovation getting harder? NBER Working Paper 11360. http://www.nber.org/papers/w11360 Last Accessed 16.05.01.

Jones, B. F. (2010). As science evolves, how can science policy? NBER Working Paper 16002. http://www.nber.org/papers/w16002 Last Accessed 16.05.01.

Koopman, R., Wang, S., Scharnhorst, A., & Englebienne, G. (2015). Ariadne's thread—interactive navigation in a world of networked information. *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15), ACM*, 1833–1838.

Lee, J. B., & Adorna, H. (2012). Link prediction in a modified heterogeneous bibliographic network. *Proceedings of the international conference on advances in social networks analysis and mining (ASONAM), IEEE*, 442–449.

Liu, X., & Croft, B. W. (2004). Cluster-based retrieval using language models. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04), ACM*, 186–193.

Liu, X., Yu, Y., Guo, C., Sun, Y., & Gao, L. (2014a). Full-text based context-rich heterogeneous network mining approach for citation recommendation. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '14)* (pp. 361–370). IEEE Press.

Liu, X., Yu, Y., Guo, C., & Sun, Y. (2014). Meta-path-based ranking with pseudo relevance feedback on heterogeneous graph for citation recommendation. *Proceedings of the 23rd ACM international conference on conference on information and knowledge management (CIKM '14), ACM*, 121–130.

Liu, X., Zhang, J., & Guo, C. (2013). Full-text citation analysis: A new method to enhance scholarly networks. *Journal of the American Society for Information Science and Technology*, *64*(9), 1852–1863.

Mai, J.-E. (2001). Semiotics and indexing: an analysis of the subject indexing process. *Journal of Documentation*, *57*(5), 591–622.

Maloney, C., Sequeira, E., Kelly, C., Orris, R., & Beck, J. (2013). *The NCBI handbook [Internet]* (2nd edition). Bethesda, MD: National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/books/NBK153388 Last Accessed 15.09.15

Mann, G. S., Mimno, D., & McCallum, A. (2006). Bibliometric impact measures leveraging topic analysis. *Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries (JCDL '06), ACM*, 65–74.
Mayr, P., & Scharnhorst, A. (2015). Scientometrics and information retrieval: weak-links revitalized. *Scientometrics*, *102*, 2193–2199.
Mell, P., & Grance, T. (2011). *Special Publication 800-145 the nist definition of cloud computing. Technical report*. National Institute of Standards and Technology U.S. Department of Commerce. http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf Last Accessed 16.05.01.
Mörchen, F., Dejori, M., Fradkin, D., Etienne, J., Wachmann, B., & Bundschus, M. (2008). Anticipating annotations and emerging trends in biomedical literature. *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '08), ACM*, 954–962.
Mutschke, P., & Mayr, P. (2015). Science models for search: a study of combining scholarly information retrieval and scientometrics. *Scientometrics*, *102*, 2323–2345.
Nelson, S. J. (2009). Medical terminologies that work: the example of MeSH. *Proceedings of the 10th international symposium on pervasive systems, algorithms, and networks (ISPAN 2009)*, 380–384.
Newman, D., Karimi, S., & Cavedon, L.(2009). Using Topic Models to Interpret MEDLINE's Medical Subject Headings. In *Proceedings of the 22nd Australasian Joint Conference on Advances in Artificial Intelligence (AI '09)* (pp. 270-279). Springer-Verlag.
Price, D. J. d. S. (1961). *Science since Babylon*. New Haven: Yale University Press.
Price, D. J. d. S. (1963). *Little science, big science*. New York: Columbia University Press.
Price, D. J. d. S. (1965). Networks of scientific papers. *Science*, *149*(3683), 510–515.
Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504.
Shi, C., Kong, X., Yu, P. S., Xie, S., & Wu, B. (2012). Relevance search in heterogeneous networks. *Proceedings of the 15th international conference on extending database technology (EDBT '12), ACM*, 180–191.
Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy*, *43*(8), 1450–1467.
Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining (KDD '04), ACM*, 306–315.
Sun, Y., & Han, J. (2012). Mining heterogeneous information networks: a structural analysis approach. *ACM SIGKDD Explorations Newsletter*, *14*(2), 20–28.
Sun, Y., Han, J., Aggarwal, C. C., & Chawla, N. V. (2012). When will it happen? Relationship prediction in heterogeneous information networks. *Proceedings of the fifth ACM international conference on web search and data mining (WSDM '12), ACM*, 663–672.
Sun, Y., Norick, B., Han, J., Yan, X., Yu, P. S., & Yu, X. (2012). Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '12), ACM*, 1348–1356.
Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T. (2011). Pathsim: meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB endowment*, *4*(11), 992–1003.
Sutcliffe, A. G., Ennis, M., & Hu, J. (2000). Evaluating the effectiveness of visual user interfaces for information retrieval. *International Journal of Human-Computer Studies*, *53*(5), 741–763.
Svenonius, Elaine. (2003). Design of controlled vocabularies. In *Encyclopedia of library and information science*. pp. 822–838. New York, NY: Marcel Dekker.
Tamine-Lechani, L., Boughanem, M., & Daoud, M. (2010). Evaluation of contextual information retrieval effectiveness: overview of issues and research. *Knowledge and Information Systems*, *24*(1), 1–34.
Ha-Thuc, V., & Srinivasan, P. (2008). Topic models and a revisit of text-related applications. *Proceedings of the 2nd PhD workshop on Information and knowledge management (PIKM '08), ACM*, 25–32.
Van Eck, N. J., & Waltman, L. (2014a). CitNetExplorer: a new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, *8*(4), 802–823.
Van Eck, N. J., & Waltman, L. (2014b). Visualizing bibliometric networks. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact: methods and practice* (pp. 285–320). Springer.
Van Raan, A. F. J. (2000). On growth, ageing, and fractal differentiation of science. *Scientometrics*, *47*(2), 347–362.
Vaquero, L. M., Rodero-Merino, L., Caceres, J., & Lindner, M. (2009). A break in the clouds: towards a cloud definition. *ACM SIGCOMM Computer Communication Review*, *39*(1), 50–55.
Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., et al. (2011). Approaches to understanding and measuring interdisciplinary scientific research: a review of the literature. *Journal of Informetrics*, *5*(1), 14–26.
Waltman, L., & Van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B*, *86*, 471.
White, H. D. (2007a). Combining bibliometrics, information retrieval, and relevance theory, Part 1: first examples of synthesis. *Journal of the American Society for Information Science and Technology*, *58*(4), 536–559.
White, H. D. (2007b). Combining bibliometrics, information retrieval, and relevance theory, Part 2: some implications for information science. *Journal of the American Society for Information Science and Technology*, *58*(4), 583–605.
White, H. D. (2015). Co-cited author retrieval and relevance theory: examples from humanities. *Scientometrics*, *102*, 2275–2299.
White, S., & Smyth, P. (2003). Algorithms for estimating relative importance in networks. *Proceedings of the ninth ACM SIGKDD int. conf. on knowledge discovery and data mining, ACM*, 266–275.
Wilson, P. (1968). *Two kinds of power: an essay on bibliographical control*. Berekeley: University of Californian Press.
Wolfram, D. (2015). The symbiotic relationship between information retrieval and informetrics. *Scientometrics*, *102*, 2201–2214.
Wu, Q., Zhang, C., Hong, Q., & Chen, L. (2014). Topic evolution based on LDA and HMM and its application in stem cell research. *Journal of Information Science*, *40*(5), 611–620.
Yan, R., Tang, J., Liu, X., Shan, D., & Li, X. (2011). Citation count prediction: learning to estimate future citations for literature. *Proceedings of the 20th ACM international conference on information and knowledge management (CIKM '11), ACM*, 1247–1252.
Yu, X., Ren, X., Sun, Y., Gu, Q., Sturt, B., Khandelwal, U., Norick, B., & Han, J. (2014). Personalized entity recommendation: a heterogeneous information network approach. *Proceedings of the 7th ACM international conference on web search and data mining (WSDM '14), ACM*, 283–292.