Community Detection: Topological vs. Topical

Ying Ding
School of Library and Information Science, Indiana University, 1320 E 10th, Bloomington, IN 47405, United States

Abstract

The evolution of the Web has promoted a growing interest in social network analysis, such as community detection. Among many different community detection approaches, there are two kinds that we want to address: one considers the graph structure of the network (topology-based community detection approach); the other one takes the textual information of the network nodes into consideration (topic-based community detection approach). This paper conducted systematic analysis of applying a topology-based community detection approach and a topic-based community detection approach to the coauthorship networks of the information retrieval area and found that: 1) Communities detected by the topology-based community detection approach tend to contain different topics within each community; and 2) Communities detected by the topic-based community detection approach tend to contain topologically-diverse sub-communities within each community. The future community detection approaches should not only emphasize the relationship between communities and topics, but also consider the dynamic changes of communities and topics.

1. Introduction

Scholarly communication forms scholarly networks: coauthorship networks, citation networks, and co-citation networks. Clusters can be detected within these scholarly networks as groups of coauthors, groups of cited or co-cited papers, authors or journals, or groups of co-occurring words. These clusters or groups are also called communities. There are two kinds of connections: social connection and similarity connection. The social connections are often real connections in the networks: friendship, coauthorship, interaction of biological entities, or communication between people. The similarity connections are derived connections which normally do not physically exist: such as the number of times two authors were co-cited together, or the number of times two words were co-occurring. Among many different community detection approaches, there are two kinds that we want to address: one considers the graph structure of the network (topology-based community detection approach); the other one takes the textual information of the network nodes into consideration (topic-based community detection approach).

Topology-based community detection researchers believe that parts of the real world can be modeled by a graph with nodes representing real world entities and edges representing real world relationships or interactions. For example, nodes in a social network can be people and edges can be friendship relations. Communities are detected based on the graph partitioning approach, which tries to minimize the number of edges between communities (Clauset, Newman, & Moore, 2004). So the nodes inside one community should have more intra-connections than inter-connections with other communities. The quality of generated clusters can be measured by the judgment of the minimization of conductance (Leskovec, Lang, Dasgupta, & Mahoney, 2008) or the maximization of modularity (Clauset, Newman, & Moore, 2004). The Girvan-Newman approach is one of the most commonly used topology-based community detection approaches (Girvan & Newman, 2002), which partitions the graph by gradually removing edges with high betweenness centralities. A similar but more scalable approach was later proposed by Clauset, Newman and Moore (2004). Other graph partitioning approaches include: Kernighan-Lin partition (Kernighan, & Lin, 1970), the spectral bisection method (Pothen, Simon, & Liou, 1990), max-flow min-cut theory (Ford & Fulkerson, 1956), and minimizing conductance cut (Leskovec, Kleinberg, & Faloutsos, 2005). Partitions can be ordered hierarchically if communities have different levels of structures.

Communities can be also detected based on topics from the content produced by the network entities, which can be papers they have published, blogs they have posted, or reviews they have written. Topic-based community detection researchers follow the principle that the more words the two entities share, the more similar these two entities are. Hierarchical clustering is a common topic-based community detection approach based on distance or similarity metrics (Newman, 2003). It first defines a distance metric between pairs of nodes based on the assigned “connection strength” to measure their similarity and then generates a tree in either a bottom-up or top-down manner. This tree describes how vertices can be grouped into communities and how these communities can be further grouped into meta-communities. The topic-modeling approach is another topic-based community detection approach, such as Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan, 2003) and its various extensions, for example, Author-Topic model (Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004), or Author-Conference-Topic Model (Tang, Zhang, Yao, Li, Zhang, & Su, 2008). LDA is a generative model that randomly generates the observable data for given hidden parameters. If the observable data are words coming from documents, it groups words into a small number of hidden topics for each document. Except assuming that the topic distribution has a Dirichlet prior, LDA is similar to probabilistic latent semantic analysis (pLSA). LDA can be viewed as a flat clustering approach to group entities together based on their similarities. The Hierarchical Latent Dirichlet Allocation (HLDA) extends the traditional LDA to a tree-based hierarchical clustering, which is similar to hierarchical clustering with more general topics as the top of the tree and more specific topics as the bottom of the tree (Blei, Griffiths, & Jordan, 2010). LDA is modularized and easy to extend. The Author-Topic model generates a cluster of authors and a cluster of words for a hidden topic. The Author-Conference-Topic model creates a cluster of authors, a cluster of conferences, and a cluster of words for a latent topic. These different clusters can be viewed as different communities that are tied to a certain topic.

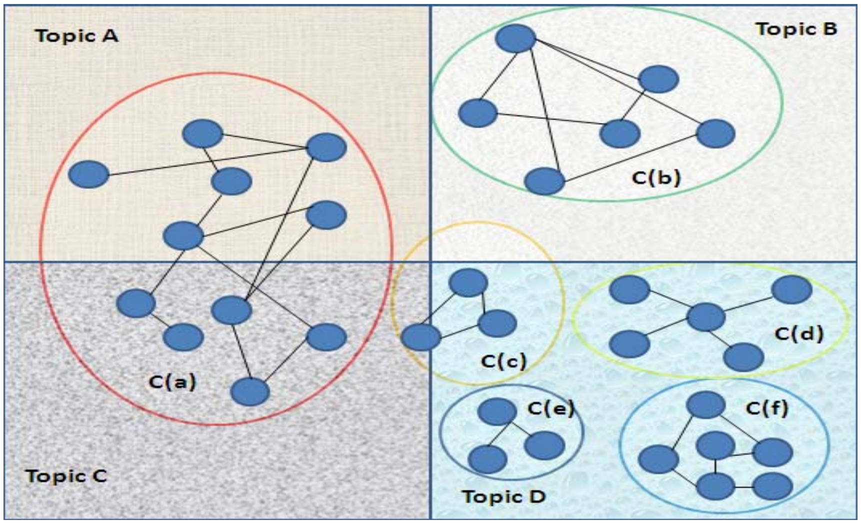


Figure 1: Communities and Topics

Topics can be highly subjective and multi-faceted, representing the subject of a discourse or a section of discourse[1]. In the current topic detection approaches, the notion of topic can be differentiated as an event-based topic or a subject-based topic. A topic that is event-based is defined to be a set of stories that are triggered by real world events. A topic that is subject-based arises out of the broader notion of subject, such as what a document is about (Allan, 2002). In this paper, the notion of topic is subject-based and can be derived from the set of scholarly publications. Since topics mainly act as the summarized "categories" of a set of documents and can have different facets, they are domain-specific, application-dependent, and context-sensitive. It is hard to justify whether a topic is "correct", but topic itself can act as a category of grouping related documents. In this paper, we are not trying to find "correct" topics, rather to find the existence of diverse topics within one community.

Communities and topics are interweaving and co-evolving (Li, He, Ding, Tang, Sugimoto, Qin, Yan & Li, 2010). A topology-based community might contain diverse topic-based sub-communities and vice versa. Figure 1 shows that Community A detected based on graph topology covers two topics (i.e., Topic A and Topic C), while the Community of Topic D contains four different sub topology-based communities (i.e., Community c, d, e, f). Taking a coauthorship network as an example, on the one hand, a topology-based community can be some research groups that group members co-authored internally with their group members but not much with external authors in the network. This kind of topology-based community can carry several topics because it is likely that one research group can collaborate with another on different research topics. On the other hand, a topic-based community can consist of different collaboration groups who collaborate under the same topic. Based on the above observations, two hypotheses can be formed:

Hypothesis 1: Communities detected by the topology-based community detection approaches tend to contain topically-diverse sub-communities within each community.

Hypothesis 2: Communities detected by the topic-based community detection approaches tend to contain topologically-diverse sub-communities within each community.

[1] Definition from Merriam-Webster Dictionary: http://www.merriam-webster.com/dictionary/topic

This paper aims to apply a topology-based community detection approach and a topic-based community detection approach to the coauthorship networks in the information retrieval (IR) field to verify whether the results are consistent with the above hypotheses. If so, then when we need to detect communities, we need to consider both topical and topological features of networks. This paper is organized as follows. Section 2 discusses the related work; Section 3 introduces the details about the dataset and the approaches used in this paper; Section 4 analyzes the results and discusses the features; and Section 5 concludes the findings and identifies future research.

2. Related Work

Topology-based Community Detection

Flake, Lawrence, Giles, & Coetzee (2002) demonstrated that link-based communities are topically related. They proposed the approximate flow cut algorithm to identify communities and tested it on the websites that are hyperlinked to the personal homepages of three prominent scientists. They found that the majority of web pages in each identified community are highly topically related in nontrivial ways. But they did not go further to analyze different topics inside each community. As the example of their Francis Crick Community, the topics can be further grouped into: Person ("crick", "francis crick", "nobel", "watson") or subject ("DNA", "biology", "genetics", "molecular"). The Table 1 in their paper explains that the topology-based community covers different topic-based sub-communities. Newman (2004) surveyed the traditional topology-based community detection algorithms and pointed out that most approaches about graph partitioning are iterative bisection: continuously dividing one group into two groups. Graph partitioning approaches bear the disadvantages that each loop has to precisely cut one group into two groups, and the number of communities which should be in a network is unknown. Therefore, traditional graph partitioning algorithms are not ideal for analyzing general network data (Newman, 2004). Girvan and Newman (2002) proposed the Girvan-Newman algorithm to extract community by gradually removing edges with high betweenness centralities in a descending order, which avoided the arbitrary bisection. The number of communities to be extracted can be measured using modularity (Newman, 2004), which measures the quality of the partitioning. Later on, Clauset, Newman and Moore (2004) proposed a scalable optimization of the Girvan-Newman algorithm by using greedy algorithm to optimize modularity. Leskovec, Lang, Dasgupta, and Mahoney (2008) introduced the concept of a Network Community Profile (NCP) plot to measure the goodness of the detected communities based on the conductance. They found that smaller communities could be combined into meaningful larger communities.

Topic-based Community Detection

Topic-based community detection approaches are based on similarity matrix or distance matrix, which edges are not real connection rather representing the similarity of two nodes. Hierarchical clustering can be used on similarity matrix to extract communities (Newman, 2004). The single linkage hierarchical clustering is a commonly used bottom-up approach that adds more and more edges to the communities based on the decreasing order of similarity. Hierarchical clustering does not require specifying the number and size of clusters to be extracted beforehand. The user can set up different thresholds to cut the dendrogram to get different sizes and numbers of communities (Newman, 2004). A similar approach has been applied in bibliometrics. The author co-citation analysis (ACA) has been widely applied to portray the intellectual structures of a domain (White & McCain, 1998), which uses hierarchical clustering to group authors based on the similarity measures of their co-citation frequencies (Ding, Chowdhury, Foo, & Qian, 2000). Latent Semantic Analysis (LSA) is a widely adopted approach to map the high dimensional co-occurrence matrix into a lower dimensional representation as latent semantic space to reveal semantic relations between entities (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). Hofmann (1999) made the significant leap forward to LSA and proposed the probabilistic LSA (pLSA) that the detected clusters are more topic-oriented. Blei, Ng, and Jordan (2003) proposed the Latent Dirichlet Allocation (LDA), a three-level hierarchical Bayesian model that models words and documents over an underlying set of topics, to avoid the pLSA's serious problems of over-fitting. Rosen-Zvi, Griffiths, Steyvers, and Smyth (2004) introduced the Author-Topic model to extend LDA to consider authorship information. Each author is associated with a multinomial distribution over topics and the clusters of authors can be detected. They compared the communities detected by the Author-Topic model with the traditional author co-word clustering and found that communities detected by the Author-Topic model have high topic similarities. Tang, Zhang, Yao, Li, Zhang, and Su (2008) extended the Author-Topic model to include publication venues so that each author is associated with a multinomial distribution over topics, words he/she wrote, and the conferences in which he/she published. Mimno, Wallach, and McCallum (2007) suggested a community-based generative model called Community-Author-Topic (CAT) that clusters both text and authors based on the notion of communities. They found that individual prolific authors could be spread through different communities reflecting the fact that they might write about different topics with different authors. Nguyen, Phung, Adams, Tran, and Venkatesh (2010) extracted communities from blogosphere based on the content of the blogs by applying LDA. Then they further clustered the words appearing in each topic to discover meta-communities. They also collated the sentiments of the blogs to each community and its meta-community. So each meta-community has sentiment and topic that can serve as a barometer to measure the mood or topic trend of these meta-communities. They found that some meta-communities contain different sentiment and topic grouping: one sentiment group can correspond to a mixture of topics and one topic group can have a collection of sentiments.

The merger of the two: topology and topics

Gruhl, Guha, Liben-Nowell, & Tomkins (2004) found that topics in the blogspace evolve due to the development of the social communities. Based on this assumption, Zhou, Ji, Zha and Giles (2006) introduced a model of the topic dynamics in social documents that connect the temporal topic dependency with the social interactions. Later on, they proposed two generative Bayesian models for semantic community detection in social networks by combining probabilistic modeling with community detection algorithms (Zhou, Manavoglu, Li, Giles, & Zha, 2006). By applying their algorithms on email corpus, they found that their approach detects the communities of individuals and in addition provides semantic topic descriptions of these communities. They defined a semantic community in a social network as including users with similar communication interests and topics that are associated with their social interactions. The proposed the Community-User-Topic model extends LDA to reflect that a community forms because its users communicate frequently and share common topics. There are two versions of this model: modeling community with users (CUT1) and modeling community with topics (CUT2). CUT1 considers a community solely as a multinomial distribution over users and relaxes the community's impact on the generated topics, which leads to a loose connection between community and topic. So the communities discovered by CUT1 are similar to the communities detected by topology-based community detection algorithms. CUT2 groups users to the same community based on shared common topics even though some of them rarely communicate. Their study found that the community identified by CUT1 contains different topics, and community identified by CUT2 contains different users coming from different departments but sharing similar topics. Li, He, Ding, Tang, Sugimoto, Qin, Yan, and Li (2010) combined LDA with the Girvan-Newman community detection algorithms using an inference mechanism and tested their algorithms on social tagging data. Their results showed that 1) users in the same community tend to be interested in similar set of topics in all time periods; and 2) topics may divide into several sub-topics and scatter into different communities over time. They found that topics seem to be the driving reasons for dynamic changes of communities: emerging, blending, and disappearing over time.

Community detection in information science

If co-citation clustering (Small, 1973; White & Griffith, 1981) and bibliographic coupling (Kessler, 1963) can be viewed as a kind of community detection effort, the history of such in bibliometrics can be traced back to the early 1970s. Bibliometricans applied different clustering approaches to identify the research fields, map the "school of thoughts", or portray the intellectual landscapes mainly based on the co-citations of authors (White & McCain, 1998), papers (Small, 1973), journals (Ding, Chowdhury, & Foo, 2000), or words (Ding, Chowdhury, & Foo, 2000a). A community in bibliometric analysis can be represented as a cluster of authors, papers, journals, or words. The major clustering approach is the hierarchical clustering or k-means clustering and a matrix can be distance or similarity based. K-means partitions nodes into clusters in which each node belongs to the cluster with the nearest mean. For example, Modha and Spangler (2000) used the toric k-means to cluster the combined similarity matrix of terms, in-links and out-links. The toric k-means uses Voronoi or Dirichlet partitions and the geometry of torus to determine the shape and the structure of the clusters. Janssens (2007) used the agglomerative clustering on an unbiased combination of textual content and citation links based on the Fisher's inverse chi-square (Janssens, Zhang, de Moor, & Glänzel, 2009). Liu, et. al. (2010) proposed the hybrid clustering algorithms including clustering ensemble which uses a consensus function to partition, and kernel-fusion clustering which clusters datasets into a high dimensional feature space and combines them as kernel matrices. The newly developed topology-based community detection approaches (i.e. the Girvan-Newman approach) are rarely deployed in bibliometrics. Recently, Wallace, Gingras and Duhon (2008) applied the extended Girvan-Newman approach to identify scientific specialties for cocitation networks. They tested this approach using raw author cocitation data from a variety of disciplines and in different time periods, and found that the results reveal the presence of distinct and identifiable scientific specialties. The advantage of Girvan-Newman approach compared with the traditional k-means clustering is that there is no need to provide the number of clusters in advance, or find the cutting point for the dendrogram if using hierarchical clustering.

Recently, textual information has been considered in bibliometric clustering. Liu, Yu, Janssens, Glanzel, Moreau, and de Moor (2010) proposed a hybrid clustering framework to incorporate lexical similarity into journal citation analysis and found that the combination of link-based clustering with textual information can improve the efficiency and usability of cocitation analysis. Other efforts of combining textual information with the cocitation clustering include Braam, Moed, and van Raan (1991), Zitt and Bassecoulard (1994), Ahlgren and Colliander (2009), and Glenisson, Glanzel, Janssens, de Moor (2005). Boyack and Klavans (2010) provided a comprehensive summary and comparison of the commonly used bibliometric clustering approaches: co-citation analysis, bibliographic coupling, direct citation, and a bibliographic coupling-based citation-text hybrid approach. They found that the hybrid approach improves upon the bibliographic coupling results in all respects. Zitt, Lelu, & Bassecoulard (2011) compared the citation-based and word-based thematic mapping approaches on the large-size document sets in the nanoscience field. The same clustering approach has been applied to the citation-based network and the word-based network separately. They found that the outcomes of these two approaches are not convergent. The efforts of hybrid clustering of text and citation either use textual information to improve similarity measures or apply traditional clustering (mostly k-means) on text matrixes and citation matrixes. Few of them applied the newly developed topology-based approaches (i.e., the Girvan-Newman approach) and further integrated textual information into the community detection approaches.

All these related works either applied topology-based community detection approaches or topic-based community detection approaches. Even though some initial efforts have realized the needs to combine both topological and topic features to detect communities, they did not systematically test the two hypotheses mentioned in this paper from the network science perspective.

3. Methodology

Datasets

Information Retrieval (IR) was selected as the field to test the proposed two hypotheses. Papers and citations from Web of Science (WOS) were collected for the latest 15 years (1993-2008). The search query contained the following terms, including their plurals or spelling variations: INFORMATION RETRIEVAL, INFORMATION STORAGE and RETRIEVAL, QUERY PROCESSING, DOCUMENT RETRIEVAL, DATA RETRIEVAL, IMAGE RETRIEVAL, TEXT RETRIEVAL, CONTENT BASED RETRIEVAL, CONTENT-BASED RETRIEVAL, DATABASE QUERY, DATABASE QUERIES, QUERY LANGUAGE, QUERY LANGUAGES, and RELEVANCE FEEDBACK. In total, 12,146 papers were collected. The whole 15-year time span was divided into two phases: Phase 1 (1993-2000) and Phase 2 (2001-2008). Table 1 shows the overview of the dataset. Phase 1 contains 3,750 articles with 9,212 authors, and its coauthorship network consists of 6,384 unique authors and 10,860 edges. Phase 2 contains 8,396 articles with 24,504 authors, and its coauthorship network consists of 13,640 unique authors and 63,140 edges.

Table 1: Overview of dataset

	1993-2000	2001-2008
No. Papers	3,750	8,396
No. Authors	9,212	24,504
Coauthor network	(6,384, 10,860)	(13,640, 63,140)

Topology-based Community Detection Approach

Girvan and Newman (2002) proposed a community detection approach (called Girvan-Newman approach) using the betweenness of the edges to identify the boundaries of communities, which measures the number of the shortest paths in a graph that use any given edge. But this approach is computationally expensive: $O(m^2n)$ on an arbitrary network with m edges and n vertices, or $O(n^3)$ on a sparse graph. This only allows the approach to be used for at most a few thousands of nodes. Later on, Clauset, Newman and Moore (2004) proposed a hierarchical agglomeration approach (called Clauset-Newman-Moore approach) to detect community that is faster than the Girvan-Newman approach. Many social networks are sparse and hierarchical and the Clauset-Newman-Moore approach can run linearly. The computational complexity is reduced to $O(mdlogn)$ where d represents the depth of the dendrogram identified in the network community structure. They have tested their approach on a customer purchasing behavior network from Amazon.com with 400,000 vertices and 2 million edges and have extracted several meaningful communities (see Figure 2).

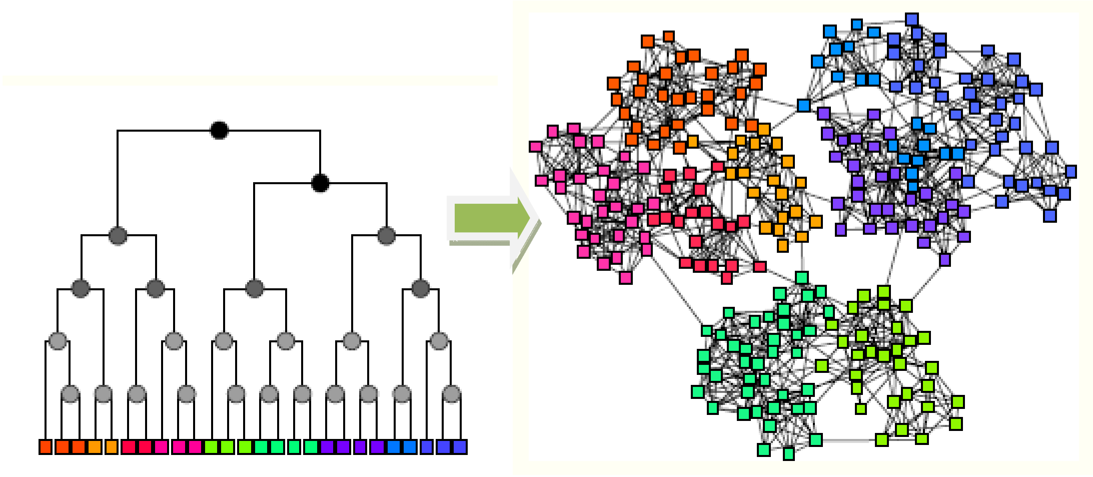


Figure 2: The Clauset-Newman-Moore approach (Clauset et al., 2008)

The Clauset-Newman-Moore approach was applied here to detect communities based on the coauthorship network of Phase 1 and Phase 2 of the dataset. Table 2 shows the detected five largest communities for each phase. The largest community in Phase 1 contains 138 authors that is 2.16% of total authors in the coauthorship network of Phase 1. The largest community in Phase 2 contains 2005 authors which is 14.7% of total authors in the coauthorship network of Phase 2. The Clauset-Newman-Moore approach tries to maximize the modularity, which often creates the issues of not being able to detect smaller clusters. Modularity has an implicit assumption about the random null model which assumes that each node can get attached to any other nodes of the network. This assumption can be problematic once the network grows larger. It can imply that the expected number of edges between two groups of nodes decreases if the size of the network increases. So, if a single edge between the two clusters is interpreted by modularity as a strong sign of the correlation, then optimizing modularity will merge these two clusters into one larger cluster. So, optimizing modularity in large networks fails to identify small clusters (Fortunato and Barthelemy, 2007).

Table 2: The five largest communities

	No. of Authors 1993-2000 (Phase 1)	No. of Authors 2001-2008 (Phase 2)
First Community	138	2005
Second Community	115	1320
Third Community	106	647
Fourth Community	90	636
Fifth Community	41	277

Topic-based Community Detection Approach

Authors can be clustered based on the similar topics they published. Rosen-Zvi, Griffiths, Steyvers, and Smith (2004) proposed the Author-Topic model to cluster both documents and authors based on their topic similarity. The Author-Topic model can extract hundreds of topics from a large-size of corpus of scientific publications. For a given topic, it is the cluster of the list of documents, the list of words, and the list of authors related to this topic.

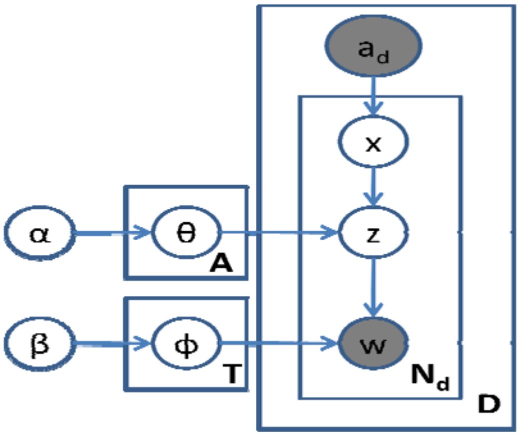


Figure 3. The Author-Topic Model

In the Author-Topic model (see Figure 3), an author is chosen randomly when a group of authors a_d decide to write a document *d* containing several topics. A word *w* is generated from a distribution of topics specific to a particular author. There are two latent variables, *z* and *x*. The formula to calculate these variables is:

$$P(z_i, x_i | z_{-i}, x_{-i}, w, a_d, \alpha, \beta) \propto \frac{C_{mj}^{wT} + \beta}{\sum_{m'}(C_{m'j}^{wT} + V\beta)} \times \frac{C_{kj}^{AT} + \alpha}{\sum_{j'}(C_{kj'}^{AT} + T\alpha)}$$

where z_i and x_i represent the assignments of the *i*th word in a document to a topic *j* and an author *k* respectively, *w* represents the observation that the *i*th word is the *m*th word in the lexicon, z_{-i} and x_{-i} represent all topic and author assignments not including the *i*th word, and C_{kj}^{AT} is the number of times an author *k* is assigned to a topic *j*, not including the current instance. The random variables ϕ (the probability of a word given a topic) and θ (the probability of a topic given an author) can be calculated as:

$$\phi_{mj} = \frac{C_{mj}^{wT} + \beta}{\sum_{m'}(C_{m'j}^{wT} + V\beta)}$$

$$\theta_{kj} = \frac{C_{kj}^{AT} + \alpha}{\sum_{j'}(C_{kj'}^{AT} + T\alpha)}$$

The outcome of this model is a list of topics, each of which is associated with the top-ranked authors, words, and documents. Top-ranked authors are not necessarily the most highly cited authors in that area, but are the productive authors who produce the most words for a given topic. The Author-Topic Model was applied here on two phases and five topics were extracted for each phase. The list of authors belong to one topic is called a community here. The number of clusters is set to be five which are based on several bibliometric mapping researches that usually there are around five major research sub-fields inside the information retrieval field (Ding, Chowdhury, and Foo, 2000). Table 3 shows the number of authors in each of the five communities that are corresponding to the five extracted topics.

Table 3: The five communities

	No. of Authors 1993-2000 (Phase 1)	No. of Authors 2001-2008 (Phase 2)
First Community	1744	3850
Second Community	1504	3773
Third Community	1340	3533
Fourth Community	1214	2670
Fifth Community	1118	2418

Note: one author can belong to more than one community

Topic Extraction Approach

Latent Dirichlet Allocation (LDA) provides a probabilistic model for the latent topic layer (Blei, Ng, & Jordan, 2003). For each document *d*, a multinomial distribution θ_d over topics is sampled from a Dirichlet distribution with parameter α. For each word w_{di}, a topic z_{di} is chosen from the topic distribution. A word w_{di} is generated from a topic-specific multinomial distribution $\phi_{z_{di}}$. The probability of generating a word *w* from a document *d* is:

$$P(w|d,\theta,\phi) = \sum_{z\in T} P(w|z,\phi_z)P(z|d,\theta_d)$$

Therefore, the likelihood of a document collection *D* is defined as:

$$P(Z,W|\Theta,\Phi) = \prod_{d\in D}\prod_{z\in T}\theta_{dz}^{n_{dz}} \times \prod_{z\in T}\prod_{v\in V}\phi_{zv}^{n_{zv}}$$

where n_{dz} is the number of times that a topic *z* has been associated with a document d, and n_{zv} is the number of times that a word w_v has been generated by a topic *z*. Comparing with Language Model (Ponte & Croft, 1998) and probabilistic latent semantic indexing (Hofman, 1999), LDA adds the latent topic layer to the model to generate more topic-centered clusters of words and documents. LDA was applied here to extract different topics from the detected communities.

4. Results/Discussions

Topology-based community contains different topics

Hypothesis 1: Communities detected by the topology-based community detection approaches tend to contain topically-diverse sub-communities within each community.

1993-2000

Table 4 shows the five largest communities detected based on the coauthorship network of 1993-2000 by the Clauset-Newman-Moore community detection approach. For each community, five topics were extracted based on LDA. Some diverse topics can be discovered within the community: largest community (databases vs. image retrieval), second community (databases vs. query), third community (user feedback vs. information retrieval), fourth community (temporal vs. database vs. query language), and fifth community (query vs. multimedia vs. mining vs. Web). The Pearson correlation coefficients were calculated to see the similarity of different topics (see Table 5). Basically Table 5 shows that there exist different topics within each community for Phase 1.

Table 4: The five topology-based communities (1993-2000)

Community	Topics	Top 10 Words
Largest Community	1	databases, framework, description, schema, expressive, datalog, logic, visualizations, taxonomic, multidimensional
	2	image, retrieval, medical, systems, content-based, query, data, relevance, feedback, evaluation
	3	information, system, environment, indexing, query, pictorial, description, database, semantic, fusion
	4	retrieval, visual, information, shape, querying, content-based, text, searching, intelligent, user
	5	image, color, retrieval, semantics, interactive, similarity, sketches, regions, knowledge, survey
Second Community	1	databases, spatial, video, content-based, access, object-oriented, temporal, system, model, image
	2	information, approach, tractable, integrated, finite, multimedia, models, interface, arithmetical, databases
	3	query, languages, systems, framework, processing, objects, integration, deductive, first-order, semantics-based
	4	query, queries, languages, functions, finitely, evaluation, aggregate, constraint, incremental, relational
	5	database, views, web, object, object-oriented, data, coupled, geographic, declustering, partitioned
Third Community	1	feedback, relevant, users, excite, query, discovering, language, generation, communicative, context
	2	information, retrieval, feedback, performance, interactive, focus, quality, analytic, selection, relevance
	3	information, interaction, relevance, study, retrieval, exploratory, modeling, review, proposal, extension
	4	retrieval, rules, implications, intermediary, learning, partial, design, windows, phrases, algorithms
	5	information, study, mediated, search, document, query, searching, systems, queries, user
Fourth Community	1	data, information, temporal, views, non-temporal, sources, warehousing, algorithms, extensions, expressiveness
	2	query, database, languages, queries, object, databases, creation, approach, temporal, transformation
	3	query, languages, language, computation, expressive, Boolean, data, kolmogorov, model, web
	4	xml, querying, graph, Euclid, expressive, analysis, completeness, engeler, model, transformation
	5	spatial, databases, extended, models, first-order, limitations, linear, database, languages, desirability
Fifth Community	1	querying, databases, non-monotonic, disjunctive, logics, programming, reasoning, empirical, semantics, consistency
	2	query, processing, image, object-based, system, evaluation, web, retrieval, class, logic
	3	database, multimedia, systems, search, hybrid, integration, semantics, hypermedia, mobility, video
	4	information, web, retrieval, models, libraries, representations, video, world-wide, query, processing
	5	data, knowledge, query, mining, rule, discovery, databases, advanced, video, phrasal

Table 5: Correlation of different topics within each community (1993-2000)

Largest Community	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic 1	1	0.029	0.109	-0.028	-0.168*
Topic 2		1	0.152*	0.166*	0.447*
Topic 3			1	0.171*	-0.127
Topic 4				1	0.251**
Topic 5					1
Second Community	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5

Topic 1	1	-0.085	0.128	-0.03	-0.043
Topic 2		1	-0.136*	-0.162*	-0.112
Topic 3			1	0.373**	-0.158*
Topic 4				1	-0.139*
Topic 5					1
Third Community	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic 1	1	-0.1	-0.145	-0.215**	-0.139
Topic 2		1	0.442**	0.139	0.312**
Topic 3			1	0.008	0.345**
Topic 4				1	-0.196*
Topic 5					1
Fourth Community	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic 1	1	-0.036	-0.094	-0.011	-0.114
Topic 2		1	0.325**	0.074	0.098
Topic 3			1	-0.137	-0.106
Topic 4				1	-0.038
Topic 5					1
Fifth Community	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic 1	1	-0.084	-0.07	-0.165*	-0.066
Topic 2		1	-0.131	0.044	-0.009
Topic 3			1	-0.111	-0.071
Topic 4				1	-0.166*
Topic 5					1

Notes: *. Correlation is significant at the 0.05 level (2-tailed); **. Correlation is significant at the 0.01 level (2-tailed).

2001-2008

Table 6 shows the five largest communities detected based on the coauthorship network of 2001-2008 by the Clauset-Newman-Moore community detection approach. For each community, five topics were extracted based on LDA. Some diverse topics can be discovered within the community: largest community (query processing vs. image retrieval), second community (information retrieval vs. semantic web), third community (database vs. description logic), fourth community (multimedia retrieval vs. cross-language retrieval), and fifth community (image retrieval vs. document retrieval). The Pearson correlation coefficients were calculated to test the similarity of different topics (see Table 7). Actually, there are many significantly correlated topics within one community at the confidence level of 0.01. but most of them have very low correlations. The fourth and fifth communities do contain different topics inside their own communities. So Table 7 can be interpreted that there do exist different topics or low-correlated topics within each community. Using the Phase 1 and Phase 2 data, we found that communities detected by the topology-based community detection approach tend to contain different topics within each community.

Table 6: The five topology-based communities (2001-2008)

Topology-based Community	Topics	Top 10 Words
Largest Community	1	image, system, retrieval, database, digital, information, analysis, video, semantic, algorithm
	2	retrieval, search, query, web, information, study, relevance, document, analysis, systems
	3	retrieval, image, feedback, relevance, content-based, learning, color, visual, images, feature
	4	query, processing, xml, data, efficient, queries, databases, approach, spatial, index
	5	retrieval, information, method, model, document, fuzzy, approach, image, web, text
Second	1	retrieval, information, query, model, document, web, feedback, relevance, documents, expansion

Community	2	retrieval, information, text, clef, experiments, system, approach, cross-language, term, medical
	3	query, processing, xml, data, efficient, queries, information, mobile, spatial, networks
	4	semantic, data, web, retrieval, similarity, system, video, multimedia, search, objects
	5	images, retrieval, system, content-based, feature, color, visual, shape, document, multimedia
Third Community	1	database, query, queries, databases, efficient, xml, data, objects, processing, networks
	2	xml, data, query, xquery, language, complex, databases, topological, visual, semistructured
	3	query, languages, relational, databases, spatial, queries, database, calculus, expressive, algebra
	4	data, querying, temporal, framework, sources, retrieval, semantic, information, model, performance
	5	description, semantic, logic, retrieval, probabilistic, logics, information, approach, fuzzy, context
Fourth Community	1	retrieval, image, medical, data, system, content-based, access, similarity, design, collections
	2	text, information, search, medline, biomedical, health, automatic, systems, analysis, searching
	3	retrieval, information, evaluation, interactive, web, relevance, effectiveness, search, video, automatic
	4	information, retrieval, query, geographic, expansion, speech, overview, spatial, digital, distributed
	5	retrieval, cross-language, clef, document, image, track, spoken, information, query, translation
Fifth Community	1	retrieval, document, model, Bayesian, structured, network, approach, web, documents, logical
	2	retrieval, image, evaluation, content-based, video, adaptive, similarity, relevance, measures, performance
	3	retrieval, information, feedback, user, search, study, interface, support, interaction, interactive
	4	retrieval, information, effectiveness, methods, user, xml, evaluating, query, interface, web
	5	information, relevance, access, framework, spoken, analysis, mobile, interpretation, modeling, concept

Table 7: Correlation of different topics within each community (2001-2008)

Largest Community	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic 1	1	0.520**	0.598**	0.235**	0.590**
Topic 2		1	0.498**	0.420**	0.638**
Topic 3			1	0.139**	0.616**
Topic 4				1	0.273**
Topic 5					1
Second Community	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic 1	1	0.711**	0.461**	0.497**	0.584**
Topic 2		1	0.374**	0.443**	0.532**
Topic 3			1	0.484**	0.306**
Topic 4				1	0.393**
Topic 5					1
Third Community	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic 1	1	0.442**	0.445**	0.298**	0.109**
Topic 2		1	0.415**	0.466**	0.196**
Topic 3			1	0.206**	0.170**
Topic 4				1	0.352**
Topic 5					1
Fourth Community	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic 1	1	0.079*	0.491**	0.444***	0.500**
Topic 2		1	0.260**	0.307**	0.067
Topic 3			1	0.607**	0.459**
Topic 4				1	0.419**
Topic 5					1
Fifth Community	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic 1	1	0.288**	0.399**	0.447**	0.07
Topic 2		1	0.506**	0.372**	0.275**
Topic 3			1	0.702**	0.486**
Topic 4				1	0.290**
Topic 5					1

Notes: *. Correlation is significant at the 0.05 level (2-tailed); **. Correlation is significant at the 0.01 level (2-tailed).

Topic-based community contains different sub-communities

Hypothesis 2: Communities detected by the topic-based community detection approaches tend to contain topologically-diverse sub-communities within each community.

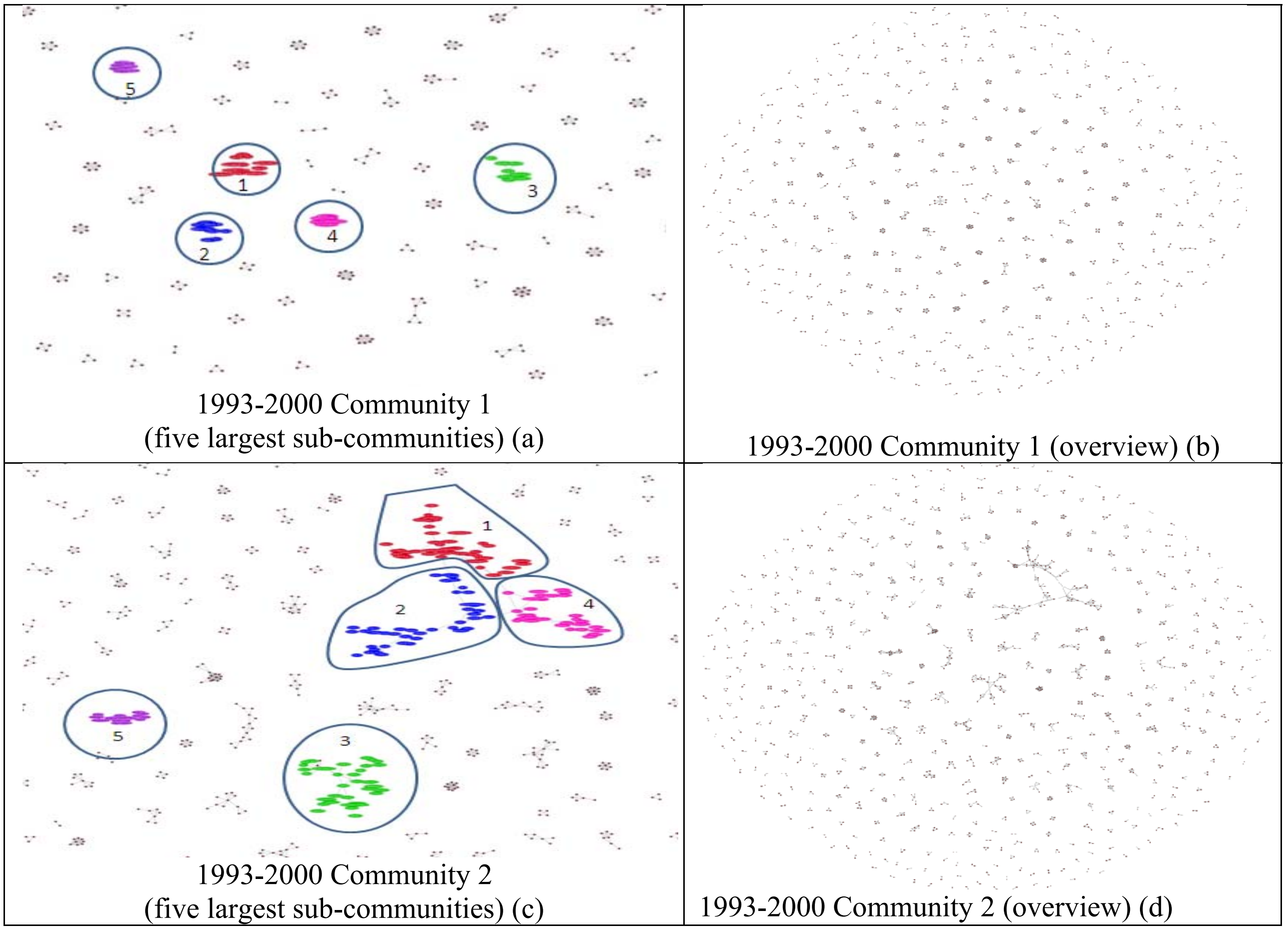
1993-2000

The Author-Topic model was applied to detect five topic-based communities for each phase. For each topic-based community, the Clauset-Newman-Moore approach was used to detected sub-communities and five largest ones were shown in Appendix I (Phase 1) and Appendix II (Phase 2). Only the top 10 highly cited authors in each sub-communities were listed as the representative authors. From the sub-communities of 1993-2000, we can figure out the research themes for these five topic-based communities: Community 1 (information science), Community 2 (database), Community 3 (information retrieval), Community 4 (medical information retrieval), and Community 5 (multimedia retrieval). Below highlights some results of 1993-2000 (see Appendix I) to demonstrate the existence of different meaningful sub-communities:

- Community 1
  - The largest sub-community is the author's co-authored colleagues: Chowdhury GG and Foo S are the author's PhD supervisors; and Liew CL and Meyyappan N are the author's PhD colleagues.
- Community 2
  - The largest sub-community captures the coauthorship network of Abiteboul S: he co-authored with Kanellakis PC on schema method, van Gucht D on contextual relations, and Gyssens M on nested relations. Besides, van Gucht D, Gyssens M, Paredanes J, van den Busscher J and Andries M also co-authored several articles on a system called GOOD.
- Community 3
  - The largest sub-community identifies the coauthorship network of Hsinchun Chen: He co-authored with Orwig RE, Nunamaker JF on a self-organizing approach to classifying electronic meeting output; with Houston A, Hubbard SM, Schatz BR on medical data mining and digital libraries; and with Houston A, Nunamaker JF and Yen J on intelligent meeting agents.
  - The second largest sub-community captures the collaboration network of Salton G;
  - The third sub-community shows the collaboration network of Roberston SE and Jones KS;
  - The fourth sub-community identifies the collaboration network of Saracevic T and Spink A;
  - The fifth sub-community is geo-oriented by grouping Korean researchers together.
- Community 4
  - The largest sub-community of Community 4 captures the collaboration about Genome Sequence Database published in Nucleic Acids Research by Harger C, Skupski M, Thompson R, Rohrlich J, Harris L, Kenn G, Easley D, and Huang W.
- Community 5

- The largest sub-community is the collaboration network of Jain AK: he co-authored with Zhang HJ, Vailaya A, Lakshmanan S, Zhong Y, and Karu K on image retrieval.
- The third sub-community captures the image retrieval research by Simth JR and Rui Y.

Figure 4 visualizes the five detected topic-based communities and their five sub-communities in Phase 1 based on the GUESS visualization system provided via the Network Workbench (NWB) at the Indiana University. The top five sub-communities were highlighted: the largest sub-community (marked with Number 1), the second sub-community (marked with Number 2), the third sub-community (marked with Number 3), the fourth sub-community (marked with Number 4), and the fifth sub-community (marked with Number 5). We can see that the overview graphs of the five topic-based communities (Figure 4 b, d, f, h, and j) are very scattered. The highlighted sub-communities in Figure 4 (a, c, e, g, i) have the good capture of the local collaboration networks and few of them are connected except the Figure 4c. It shows that in 1993-2000, information retrieval was conducted by different sub-communities and collaboration between these sub-communities is rare.

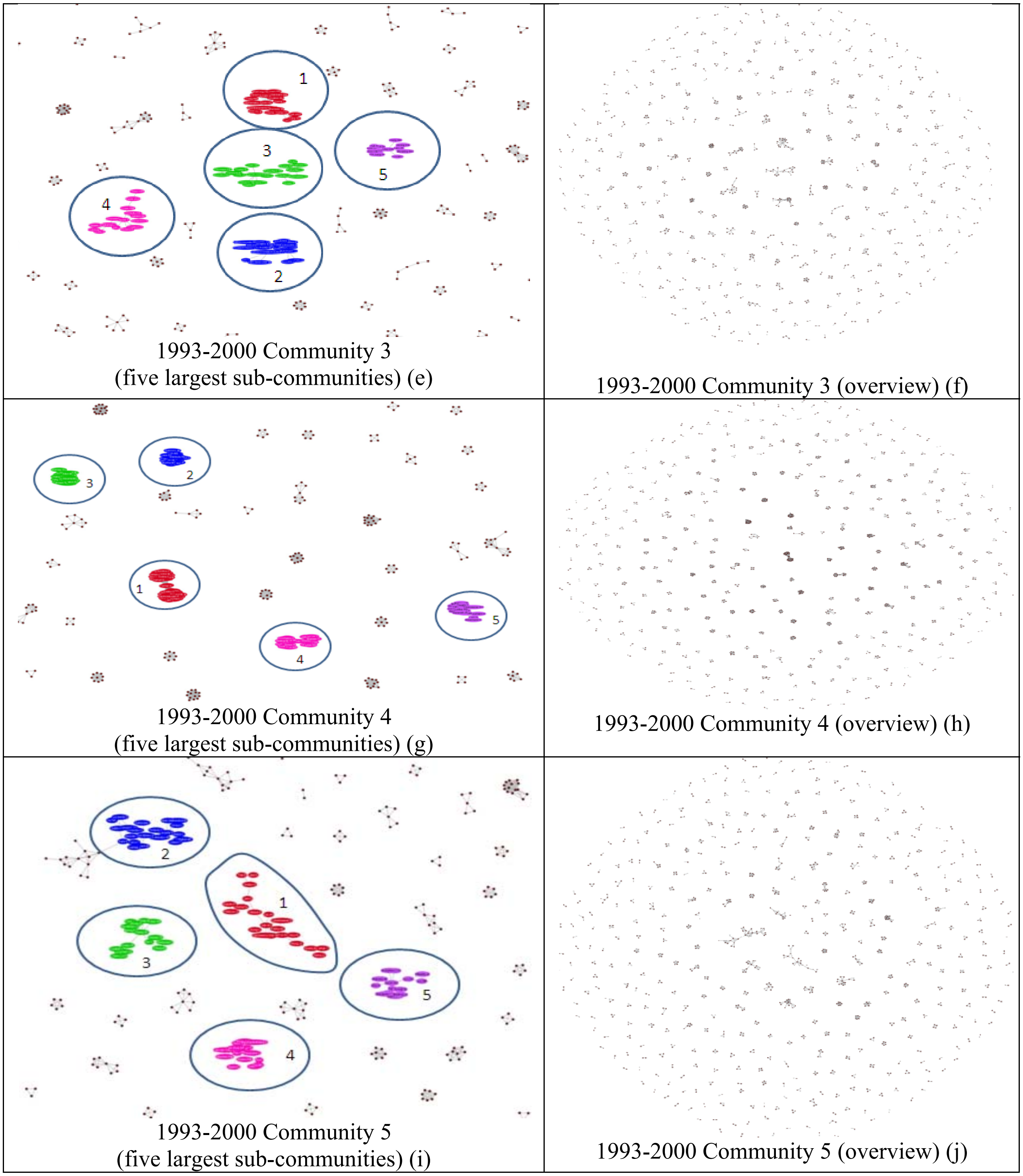


1993-2000 Community 1
(five largest sub-communities) (a)

1993-2000 Community 1 (overview) (b)

1993-2000 Community 2
(five largest sub-communities) (c)

1993-2000 Community 2 (overview) (d)



1993-2000 Community 3
(five largest sub-communities) (e)

1993-2000 Community 3 (overview) (f)

1993-2000 Community 4
(five largest sub-communities) (g)

1993-2000 Community 4 (overview) (h)

1993-2000 Community 5
(five largest sub-communities) (i)

1993-2000 Community 5 (overview) (j)

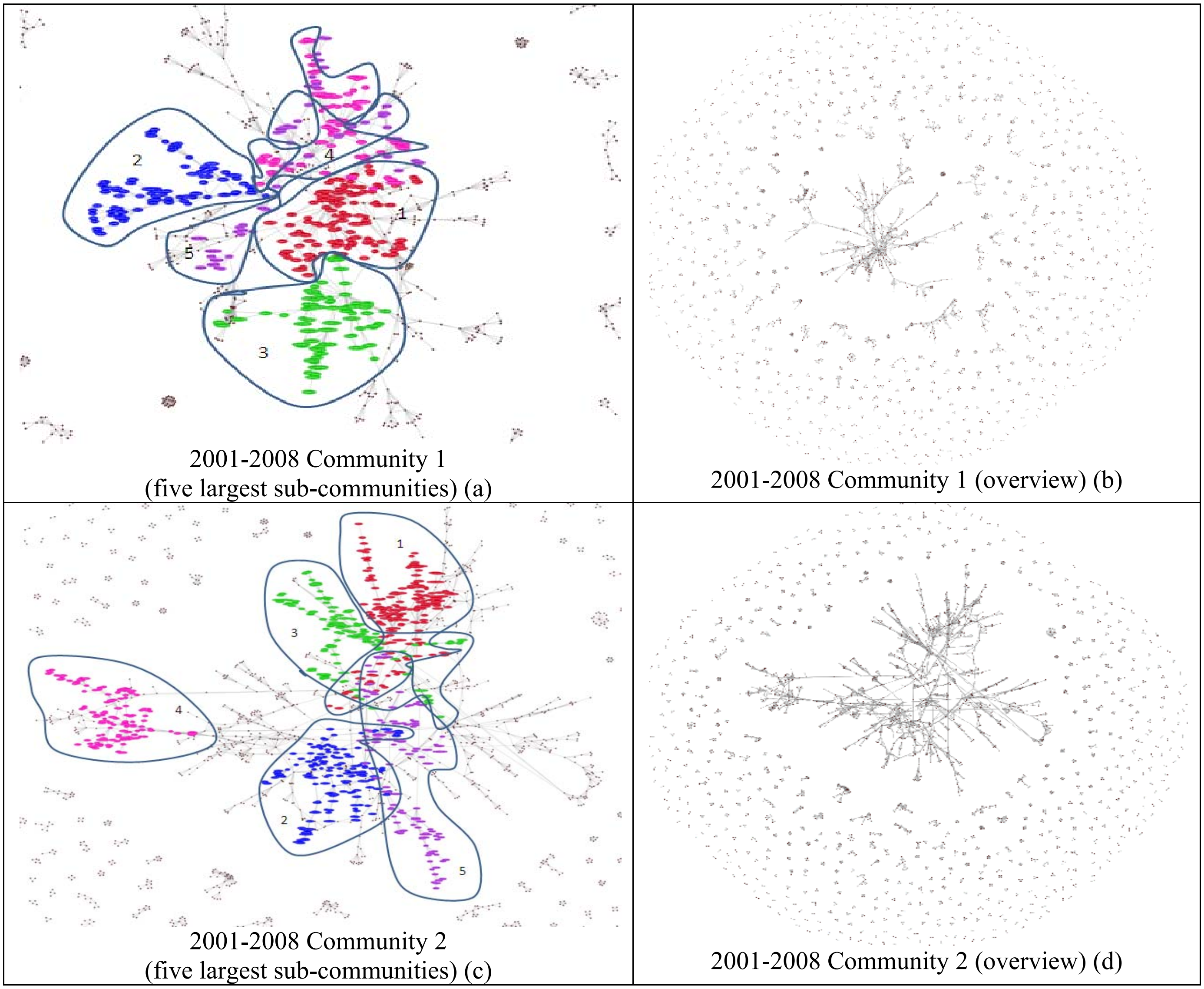
Figure 4: Topic-based communities and their sub-communities (1993-2000)

2001-2008

The Author-Topic model was applied to detect five topic-based communities for the period of 2001-2008. For each topic-based community, the Clauset-Newman-Moore approach was used to detected sub-communities and five largest ones were shown in Appendix II. Only the top 10 highly cited authors in each sub-community were listed as the representative authors. From the sub-communities of 2001-2008, we can figure out the research themes for these five topic-based communities: Community 1 (multimedia retrieval), Community 2 (database), Community 3 (medical retrieval), Community 4 (information retrieval), and Community 5 (mixture of different topics). Below highlights some results of 2001-2008 to display the identified meaningful sub-communities:

- Community 1
    - The largest sub-community shows the collaboration network of Smith JR, Jain AK and Ma WY and all of them have research focuses on image retrieval. Smith JR did not collaborate with Jain AK and Ma WY, but Jain AK and Ma WY wrote one paper together on relevance feedback for natural image retrieval.
    - The second sub-community features the collaboration network of Kittler J,
    - The third sub-community shows the network of Smeulders AW,
    - The fourth sub-community identifies the collaboration network of Rui Y. All of them are well-known for their multimedia retrieval research.
- Community 2
    - The largest sub-community features the collaboration network of XML database groups: Abiteboul S (active XML, XML data warehouse, and XML integration), Buneman P (XML query, provenance), and Fernandez M (XML Query, XML storage).
    - The second sub-community is represented by the collaboration network of graph mining: Faloutsos C (graph database and query) and Yang Y (social network mining).
    - The fourth sub-community mainly gathers Korean researchers.
- Community 3
    - The largest sub-community represents the collaboration network of medical information retrieval: Muller H (visual and text retrieval for medical documents), Hersh WR (medical image retrieval), and Gorman PN (cognition for medical informatics).
- Community 4
    - represents several collaboration networks of major information retrieval players in the information science area: Robertson SE (largest sub-community), Chen HC (second sub-community), Marchionin G (third sub-community), Spink A (fourth sub-community) and Bates MJ (fifth sub-community).
- Community 5
    - is the mix of information retrieval (largest and fifth sub-communities), image retrieval (second sub-community), video retrieval (third sub-community), and medical retrieval (fourth sub-community).

Figure 5 illustrates the five detected topic-based communities and their five sub-communities in Phase 2 using the GUESS visualization system provided by the NWB tool. The top five largest sub-communities were highlighted: the first sub-community (marked with Number 1), the second sub-community (marked with Number 2), the third sub-community (marked with Number 3), the fourth sub-community (marked with Number 4), and the fifth sub-community (marked with Number 5). The five largest sub-communities are located in the center of its graph (Figure a, c, e, g, and i). This time, the centers of the graphs are connected (Figure 5b, d, f, h, and j) which indicates that collaboration between sub-communities are increasing. However, in Phase 1, scientific collaboration in information retrieval is not that obvious that the coauthorship networks are not densely connected. This can be related to the feature of the networks and the feature of the research field. Based on the results from the two phases, we found that there exist meaningful topology-based sub-communities inside each topic-based community.

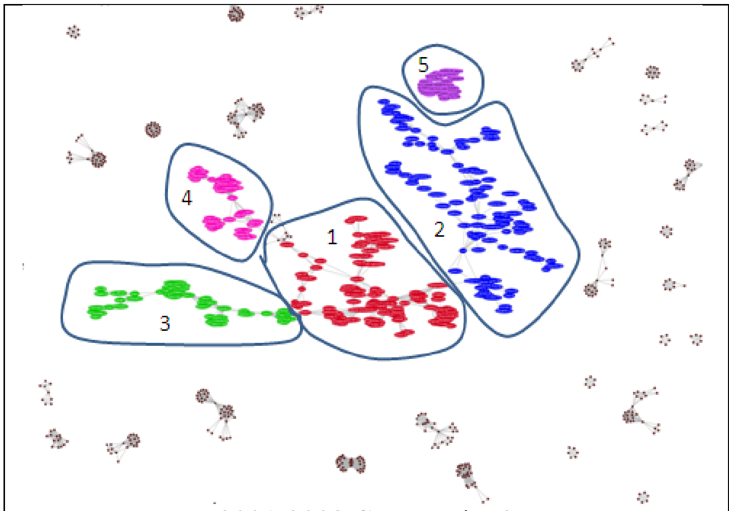


2001-2008 Community 1 (five largest sub-communities) (a)

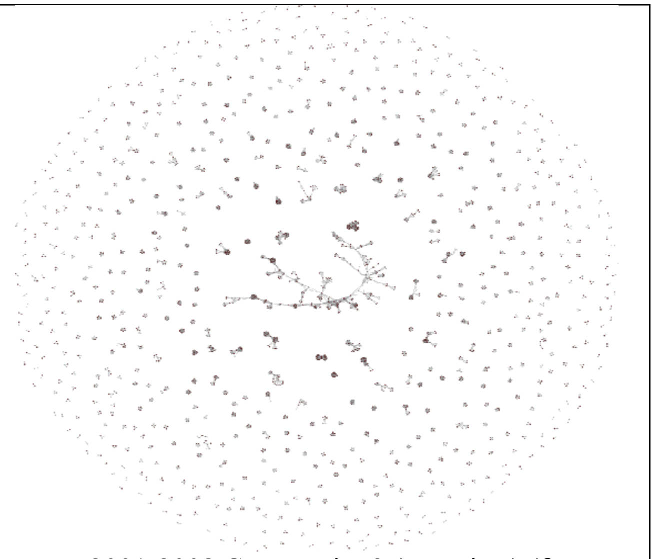
2001-2008 Community 1 (overview) (b)

2001-2008 Community 2 (five largest sub-communities) (c)

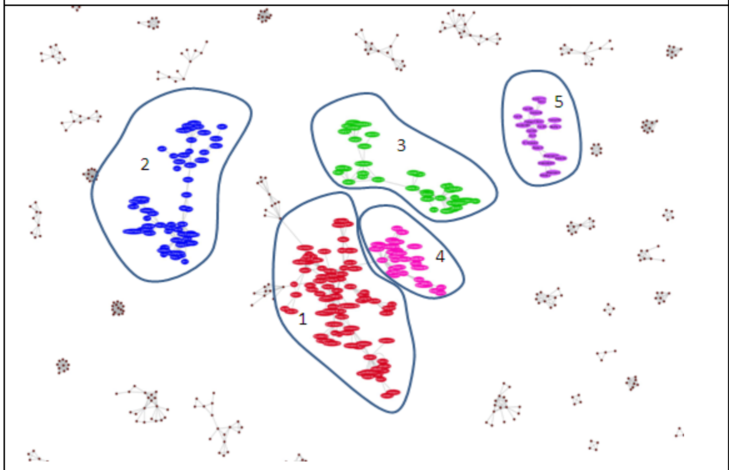
2001-2008 Community 2 (overview) (d)



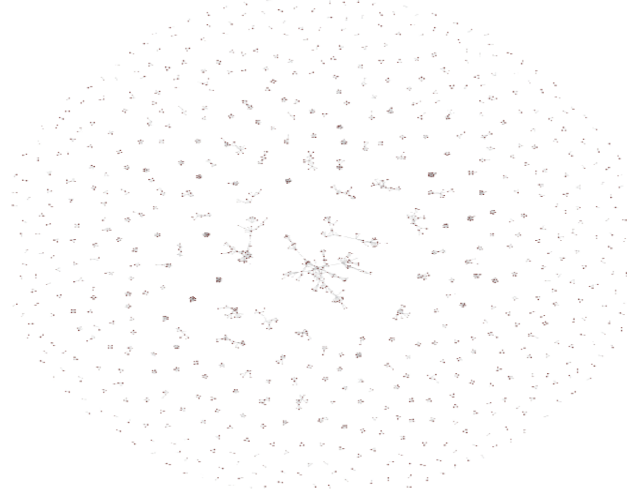
2001-2008 Community 3
(five largest sub-communities) (e)



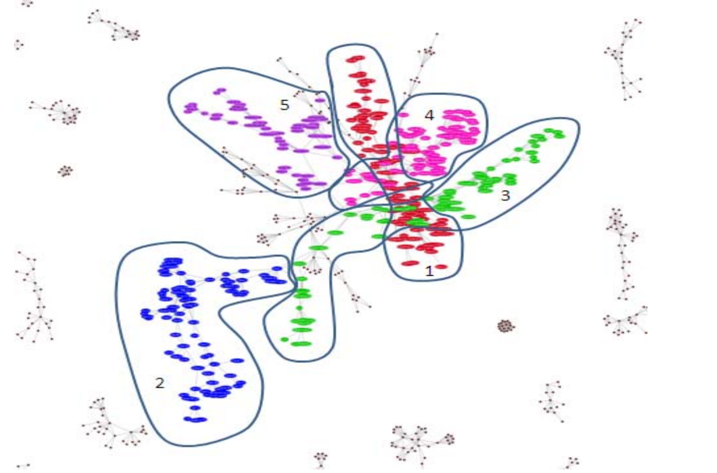
2001-2008 Community 3 (overview) (f)



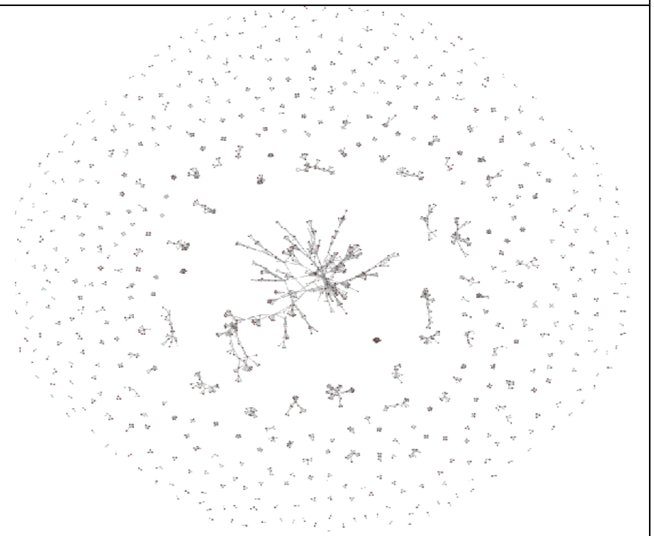
2001-2008 Community 4
(five largest sub-communities) (g)



2001-2008 Community 4 (overview) (h)



2001-2008 Community 5
(five largest sub-communities) (i)



2001-2008 Community 5 (overview) (j)

Figure 5: Topic-based communities and their sub-communities (2001-2008)

The size of communities in Phase 1 is not that large (with average 100 members), which can be considered as a good size for clusters (Leskovec, Lang, Dasgupta, and Mahoney, 2008), still the topic diversity is detected. In general, the level of clustering or partition can go repeatedly until only one node remains in the network. No matter what size of the communities being detected, there should exist the topical and topological diversity inside these communities. As mentioned in Figure 1, research groups collaborate with each other on different topics, and each topic can be researched by several different research groups.

5. Conclusion

Among many different community detection approaches, we address two kinds of approaches: topology-based and topic-base. The topology-based community detection approaches are commonly used. However, discovering a community purely based on graph topology can be problematic: 1) a spammer can generate edges to all the nodes in the networks which pose challenges to the topology-based community detection; and 2) it is hard to explain the semantic reason why such communities are formed purely based on the topology-based approach (Zhou, Manavoglu, Li, Giles, & Zha, 2006). As Wasserman and Faust (1994) pointed out, the formation of community is resulted from the similarity among social actors and purely considering topology to identify community is insufficient.

This paper conducted a systematic analysis by applying a topology-based community detection approach and a topic-based community detection approach to the coauthorship networks of the information retrieval field and the results are consistent with the following hypotheses: 1) *Hypothesis 1: Communities detected by the topology-based community detection approaches tend to contain topically-diverse sub-communities within each community*; and 2) *Hypothesis 2: Communities detected by the topic-based community detection approaches tend to contain topologically-diverse sub-communities within each community*. This implies that community detection should consider both the topological and topical features of the networks. There are some initial efforts towards such direction: Zhou, Manavoglu, Li, Giles and Zha (2006)'s Community-User-Topic model; Li, He, Ding, Tang, Sugimoto, Qin, Yan, and Li (2010)'s inference model of combining LDA and Girvan-Newman approach; and Li et al. (2011)'s semi-supervised dynamic community topic model. These early efforts mainly extended LDA by modeling communities as a multinomial distribution over the networks of authors and their topics. But how to extend the graph partitioning approach to consider the topic features is not well explored.

In bibliometrics, traditional methods have limits on the size of the networks, and the disadvantage of providing the clear-cut of the dendrogram which challenge the robustness of these approaches. Few studies have utilized topology-based community detection approaches on scholarly networks, especially on large-scale cocitation networks with more than 100 million nodes and 1 billion edges (Wallace, Gingras, & Duhon, 2008). Recent studies of combining text and citation in various ways primarily consider textual information as a kind of similarity measure. The textual information is not being fully integrated into the community detection approaches instead it acts as an input for such approaches (Liu, Yu, Janssens, Glanzel, Moreau, and de Moor, 2010). Furthermore, hybrid approaches tend to have higher computational costs (Boyack & Klavans, 2010). Clearly, there is a need to integrate topic-based and topology-based community detection approaches in bibliometrics.

This study has several limitations that show potential future research. First, the two hypotheses are not logically proven, rather than the results in this paper are found consistent with the two hypotheses. It is strongly aligned with the objective of this paper that aims to demonstrate that there exists either topological diversity when purely using a topic-based approach or topical diversity when merely applying a topology-based approach. Future study will test different topology-based and topic-based community detection approaches on different datasets to further prove these two hypotheses. Second, the datasets used in this study come from only one research field – the information retrieval (IR) field. This field is multi-disciplinarily driven and therefore could already bring the topical or topological diversity in advance. In the future, we would choose one field which is not multi-disciplinary, such as physics, or pure mathematics. We also want to choose fields covering science, social science and humanities. Third, the evaluation in this study is qualitative because it analyzes the results based on the scholarly behaviors of IR researchers. In the future, we would conduct the quantitative evaluation based on the test of different topic-based and topology-based approaches on different datasets from different research fields.

Communities are dynamic. Users freely join and leave the communities resulting in changes of community structures. Topics are dynamic. The topic focus of the communities can vary from time to time based on the current research focuses of authors. Most of the current social network analysis is focusing on describing the features of the static network. However, social networks are changing and evolving over time. The dynamic changes of community structures can greatly impact the content evolution of social or scholarly communication. Periodically clustering data and examining extracted topics can provide a snapshot of these dynamic changes. The identified dynamic patterns can be used to predict future interactions or shift of topics. Using topics to understand the dynamics of community structures and interpreting topic transformation based on the evolving social interactions are mutually important for us to better understand communities and topics. Future community detection approaches should not only emphasize the relation between community and topics, but also consider the dynamic changes of communities and topics.

Acknowledgement

The author is grateful to Prof. Jie Tang at Tsinghua University (China) for sharing the codes of several algorithms applied here, and Prof. Katy Börner at Indiana University for her team of support for the visualization part. The author is also thankful to the two anonymous reviewers for their insightful comments.

6. References

Ahlgren, P., & Colliander, C. (2009). Textual content, cited references, similarity order, and clustering: An experimental study in the context of science mapping. *Scientometrics*, 83(3), 862–873.

Allan, J. (2002). *Topic detection and tracking: Event-based information organization*. Kluwer Academic Publishers.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of ACM*, 57(2), 1-30.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1033.

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.

Braam, R.R., Moed, H.F., &Van Raan, A.F.J. (1991). Mapping of science by combined cocitation and word analysis, Part 1: Structural aspects. *Journal of the American Society for Information Science*, 42(4), 233–251.

Clauset, A., Moore, C., & Newman, M. (2008). Hierarchical structure and prediction of missing links in networks. *Nature*, 453, 98-101.

Clauset, A., Newman, M., & Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E*, 70, 066111.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.

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.

Ding, Y.**,**Chowdhury, G., & Foo, S. (2000a). Incorporating the Results of Co-word Analyses to Increase Search Variety for Information Retrieval. *Journal of Information Science*, 26(6), 429-452.

Ding, Y., Chowdhury, G., Foo, S., & Qian, W. (2000). Bibliometric Information Retrieval System (BIRS): A Web Search Interface Utilizing Bibliometric Research Results. *Journal of the American Society for Information Science*, 51(13), 1190-1204.

Flake, G. W., Lawrence, S., Giles, L. C., & Coetzee, F. M. (2002). Self-organization and identification of web communities. *IEEE Computer*, 35, 66-70.

Ford, L. R., & Fulkerson, D. R. (1956). Maximal flow through a network. *Canadian Journal of Mathematics*, 8, 399-404.

Fortunato, S., & Barthelemy, M. (2007). Resolution limit in community detection. PNAS, 104(1), 36-41.

Girvan, M., & Newman, M. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99, 7821-7826.

Glenisson, P., Glänzel,W., Janssens, F., & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, 41(6), 1548–1572.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web*, p491-501, May 17-22, 2004, New York, NY, USA.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p50-57, August 15-19, 1999, Berkeley, CA, USA.

Janssens, F., Zhang, L., de Moor, B., & Glänzel, W. (2009) Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management*, (6): 683-702

Janssens, F. (2007). *Clustering of scientific fields by integrating text mining and bibliometrics*. Doctoral Dissertation. Faculty of Engineering, Katholieke Universiteit Leuven, Belgium.

Kernighan, B. W., & Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. *Bell Systems Technical Journal*, 49, 291-307.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 24, 123-131.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). Graphs over time: Densification laws, Shrinking diameters and possible explanations. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, p177-187, August 21-24, 2005, Chicago, IL, USA.

Leskovec, J., Lang, K. J., Dasgupta, A., & Mahoney, M.W. (2008). Statistical properties of community streucture in large social and information networks. In *Proceedings of the 17th International Conference on World Wide Web*, p695-704, April 21-25, 2008, Beijing, China.

Li, D., He, B., Ding, Y., Tang, J., Sugimoto, C., Qin, Z., Yan, E., & Li, J. (2010). Community-based topic modeling for social tagging. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM2010)*, p1565-1568, October 26-30, 2010, Toronto, Canada.

Li, D., Zhu, J. Ding, Y., Xin, S., Chen, S., Tang, J., Bollen, J., & Rocha, G. (2011). Adding community and dynamics to topic models. Technical Report. School of Library and Information Science, Indiana University.

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.

Modha, D.S., & Spangler, W.S. (2000). Clustering hypertext with applications to web searching. In *Proceedings of the 7th ACM on Hypertext and Hypermedia*, p143-152, May 30 - June 3, 2000, San Antonio, TX, USA

Mimno, D., Wallach, H., & McCallum, A. (2007). Community-based link prediction with text. In *Workshop on statistical models of networks, the 21st Annual Conference on Neural Information Processing Systems (NIPS'07)*, December 3-7, 2007, Vancouver, BC, Canada.

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

Newman, M. (2004). Detecting community structure in networks. *European Physics Journal B*., 38, 321-330.

Nguyen T., Phung, D., Adams, B., Tran, T., & Venkatesh, S. (2010). Hyper-community detection in the blogsphere. In *Workshop on Social Media, the second ACM SIG Multimedia*, October 25-29, 2010, Firenze, Italy.

Ponte, J. M., & Croft, W.B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p275-281, August 24-28, 1998, Melbourne, Australia.

Pothen, A., Simon, H., & Liou, K.P. (1990). Partitioning sparse matrics with eigenvectors of graphs. *SIAM Journal of Mathematic Analysis Application*, 11(3), 430-452.

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*, p487-494, July 7-11, 2004, Banff, Canada.

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

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008)*. p990-998, August 24-27, 2008, Las Vegas, NV, USA.

Wallace, M., Gingras, Y., & Duhon, R. (2008). A new approach for detecting scientific specialties from raw cocitation networks. Available at: http://arxiv.org/abs/0807.4903 (accessed: Jan 25, 2011)

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*, Cambridge University Press.

White, H.D., & Griffith, B.C. (1981). Author co-citation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32, 163-172.

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 and Technology*, 49(4), 327-355.

Zhou, D., Ji, X., Zha, H., & Giles, L. C. (2006). Topic evolution and social interactions: How authors effect research. In *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management*, p248-257, November 6-11, 2006, Arlington, VA, USA.

Zhou, D., Manavoglu, E., Li, J., Giles, L. C., & Zha, H. (2006). Probabilistic models for discovering e-communities. In *Proceedings of the 15th ACM International Conference on World Wide Web*, p173-182, May 23-26, 2006, Edinburgh, Scotland.

Zitt, M., & Bassecoulard, E. (1994). Development of a method for detection and trend analysis of research fronts built by lexical or cocitation analysis. *Scientometrics*, 30(1), 333–351.

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.

Appendix

Appendix I

Appendix I: The five topic-based communities (1993-2000)

Topic-based Community	Sub-community	Representative authors
Community 1	Largest sub-community (1)	Chowdhury GG, Ding Y, Chowdhury S, Liew CL, Hui SC, Foo S, Lim HK, Hui L, Meyyappan N, Chennupati KR
	Second sub-community (2)	Hardin JB, Cole TW, Bishop AP, Schatz B, Chen HC, Yang CC, Mischo WH, Yen J, Zhu B,
	Third sub-community (3)	Wilson R, Landoni M, Sweeney N, Gibb F, Leon R, O'Donnell R, McCartan C, Bell S
	Fourth sub-community (4)	Neligon C, Chouinard E, Way D, Danchak M, Bachiochi D, Conlan N, Berstene M, Furey T
	Fifth sub-community (5)	Wegner R, Muller I, Schafer J, Santo H, Kaeber J, Jungblut H, Jonas K, Kaul M
Community 2	Largest sub-community (1)	Abiteboul S, Ceri S, Paredaens J, Gyssens M, Kanellakis PC, Van den Bussche J, Schurr A, Andries M, Engels G, Van Gucht D
	Second sub-community (2)	Hull R, Bertino E, Grumbach S, Libkin L, Benedikt M, Ciaccia P, Colby LS, Zezula P, Scholl M, Ooi BC
	Third sub-community (3)	Vardi MY, Kolaitis PF, Gottlob G, Eiter T, Adali S, Bell C, Subrahmanian VS, Cadoli M, Leone N, Ng RT
	Fourth sub-community (4)	Buneman P, Valduriez P, Papadimitriou CH, Suciu D, Gravano L, Florescu D, Fernandez M, Tomasic A, Raschid L, Deutsch A
	Fifth sub-community (5)	Sheth AP, Woelk D, Mena E, Kashyap V, Huhns MN, Singh MP, Tomlinson C, Perry B, Cannata PE, Nodine M
Community 3	Largest sub-community (1)	Chen HC, Shatz BR, Orwig RE, Nunamaker JF, Houston A, Zhang Y, Yen J, Ramsey M, Tolle KM, Hubbard SM
	Second sub-community (2)	Salton G, Roberston SE, Harman D, Jones KS, Buckley C, Hancockbeaulieu M, Walker S, Allan J, Lewis D, Smeaton A
	Third sub-community (3)	Robertson SE, Lewis DD, Walker S, Jones KS, Beaulieu M, Rasmussen E, Jones GJF, Young SJ, Foote JT, Cercone N
	Fourth sub-	Saracevic T, Spink A, Losee RM, Tibbo HR, Jansen BJ, Robins D, Bateman J, Goodrum

	community (4)	A, Qin J, Paris LAH
	Fifth sub-community (5)	Myaeng SH, Kang HK, Jung H, Lee JS, Yuh S, Kim YK, Jeong KS, Choi KS, Kim MC, Sim CM
Community 4	Largest sub-community (1)	Harger C, Skupski MP, Fields C, Huang W, Thompson R, Rohrlich J, Harris L, Kenn G, Easley D, Harpold M
	Second sub-community (2)	Lowe DG, Roth ME, Shenoy SG, Yang RH, Jin HK, Shimkets RA, Hillan K, Murtha MT, Went GT, Predki PF
	Third sub-community (3)	Rinsland CP, Toon GC, Gunson MR, Zander R, Chang AY, Stiller GP, Brown LR, Abrams MC, Allen M, Manney GL
	Fourth sub-community (4)	Schuler GD, Boguski MS, Tatusova TA, Madden TL, Wheeler DL, Ermolaeva O, Leipe DD, Rapp BA, Simon R, Pruitt KD
	Fifth sub-community (5)	Barker WC, Mewes HW, Wu C, Srinivasarao GY, Heumann K, Orcutt BC, Tsugita A, Ledley RS, Marzec CR, Pfeiffer F
Community 5	Largest sub-community (1)	Jain AK, Zhang HJ, Aigrain P, Swets DL, Healey G, Jain A, Vailaya A, Lakshmanan S, Zhong Y, Karu K
	Second sub-community (2)	Catarci T, Delbimbo A, Batini C, Lucarella D, Corridoni JM, DeMarsicoi M, Cinque L, Berretti S, Assfalg J, Colombo C
	Third subCommunity (3)	Smith JR, Rui Y, Chang SF, Chen W, Ortega M, Zhong D, Mehrotra S, Huang TS, Li CS, Beigi M
	Fourth sub-community (4)	Vetterli M, Squire DM, Pun T, Muller H, Squire D, Giess C, Muller W, Van der Veer GC, Pecenovic Z, Pu P
	Fifth sub-community (5)	Chen HC, Schatz B, Nunamaker JF, Ramsey M, Ng T, Lin CT, Liu DR, Zhu B, Orwig R, Lin CH

Appendix II

Appendix II: The five topic-based communities (2001-2008)

Topic-based Community	Sub-community	Representative authors
Community 1	Largest sub-community (1)	Smith JR, Jain AK, Ma WY, Xu J, Vailaya A, Wu Y, Zhang DS, Jing F, Chang E, Zhang HJ
	Second sub-community (2)	Kittler J, Lee J, Kim H, Lee S, Matas J, Park J, Lee Y, Kim KS, Lee M, Park SH
	Third sub-community (3)	Smeulders AW, Zhou XS, Gevers T, Petrakis EGM, Naphade MR, Sebe N, Tian Q, Moghaddam B, Hollink L, Huang TS
	Fourth sub-community (4)	Rui Y, Wang JZ, Li J, Yang J, Wiederhold G, Chen L, Krovetz R, Chen YX, Liu L, Kankanhalli MS
	Fifth sub-community (5)	Lee JH, Barnard K, Raghavan VV, Park S, Kim M, Chang YC, Chung CW, Kim DH, Lee KM, Kim SH
Community 2	Largest sub-community (1)	Abiteboul S, Buneman P, Fernandez M, Guting RH, Milo T, Chomicki J, Benedikt M, Theodoridis Y, Jensen CS, Tatarinov I
	Second sub-community (2)	Faloutsos C, Yang Y, Zhang J, Papadias D, Li Q, Chen H, Yang J, Chen J, Ioannidis YE, Lin D
	Third sub-community (3)	Florescu D, Madden S, Ceri S, Hellerstein JM, Stoica I, Carey MJ, Schmidt A, Hristidis V, Westerveld T, Ramakrishnan R
	Fourth sub-community (4)	Lee D, Lee J, Kim J, Kim M, Lee S, Park J, Whang KY, Lee Y, Kim KS, Lee JY
	Fifth sub-community (5)	Gupta A, Santini S, Zhang C, Jagadish HV, Chen Y, Wolfson O, Yu C, Lakshmanan LVS, Jain R, Chen YX
Community 3	Largest sub-community (1)	Muller H, Hersh W, Gorman PN, Lowe HJ, Zweigenbaum P, Yang JJ, Strauss A, Ruch P, Darmoni SJ, Lovis C
	Second sub-community (2)	Chen Y, Friedman C, Kim W, Cimino JJ, Yu H, Ely JW, Aronson AR, Soergel D, Rindflesch TC, Feldman R
	Third sub-community (3)	Stapley BJ, Paton NW, Hermjakob H, Kumar A, Nenadic G, Etzold T, Apweiler R, Kumar V, Ananiadou S, Spasic I
	Fourth sub-community (4)	Datta A, Musen MA, Berrios DC, Shah M, McQueen J, Blum M, Tu SW, Chan A, Tang Y, Mathur R
	Fifth sub-community (5)	Suzuki M, Nakajima H, Sasano Y, Yokota T, Sugita T, Kobayashi H, Irie H, Saitoh N, Kanzawa H, Hopfner M

Community 4	Largest sub-community (1)	Robertson SE, Belkin NJ, Saracevic T, Croft WB, Ingwersen P, Allan J, Vakkari P, Jarvelin K, Mizzaro S, Kekalainen J
	Second sub-community (2)	Chen HC, Chen H, Yang CC, Menczer F, Chau M, Lam W, Roussinov D, Wei CP, Xi W, Sheng ORL
	Third sub-community (3)	Marchionin G, Zhang J, Jorgensen C, Downie JS, Soergel D, Wolfram D, Rasmussen EM, Choi Y, Komlodi A, Lee ML
	Fourth sub-community (4)	Spink A, Jansen BJ, Ellis D, Wilson TD, Ford N, Ozmutlu S, Miller D, Greisdorf H, Goodrum A, Ozmutlu HC
	Fifth sub-community (5)	Bates MJ, Fox EA, Gordon M, Fan W, Gordon MD, Fan WG, Radev D, Bhavnani SK, Wu H, Pathak P
Community 5	Largest sub-community (1)	Fuhr N, Marchionini G, Crestani F, Borlund P, Ruthven I, White RW, Amati G, Lalmas M, Tombros A, Kazai G
	Second sub-community (2)	Ma WY, Xu J, Zobel, J, Nie JY, Kraaij W, Lin X, Zhang Y, Chen SM, Chua TS, Wen JR
	Third sub-community (3)	Smeaton AF, Moffat A, Braschler M, Peters C, Li Y, Agosti M, Cristianini N, Jones GJF, Anh VN, Smyth B
	Fourth sub-community (4)	Muller H, Hersh W, Oard DW, Resnik P, Clough P, Sanderson M, Beaulieu M, Lehmann TM, Larson RR, French JC
	Fifth sub-community (5)	Roberston SE, Hawking D, Savoy J, Hiemstra D, Craswell N, Wang J, Jones SP, Wang Y, Westerveld T, Bailey P