Mining diversity subgraph in multidisciplinary scientific collaboration networks: A meso perspective

Bing He[1], Ying Ding[1], Jie Tang[3], Vignesh Reguramalingam[2,] Johan Bollen[2]
[1]School of Library and Information Science, Indiana University, Bloomington, IN, USA
[2]School of Informatics and Computing, Indiana University, Bloomington, IN, USA
[3]Department of Computer Science and Technology, Tsinghua University, Beijing, China
{binghe;dingying;vignregu; jbollen}@indiana.edu, jery.tang@gmail.com

Abstract

This paper proposes a framework to analyze the interdisciplinary collaboration in a coauthorship network from a meso perspective using topic modeling: (1) a customized topic model is developed to capture and formalize the interdisciplinary feature; and (2) the two algorithms Diversity Subgraph Extraction (DSE) and Constraint-based Diversity Subgraph Extraction (CDSE) are designed and implemented to extract a meso view, i.e. a diversity subgraph of the interdisciplinary collaboration. The proposed framework is demonstrated using a coauthorship network in the field of computer science. A comparison between DSE and Breadth First Search (BSF)-based subgraph extraction favors DSE in capturing the diversity in interdisciplinary collaboration. Potential possibilities for studying various research topics based on the proposed framework of analysis are discussed.

1 Introduction

Interdisciplinary collaboration which integrates theories and methodologies from different knowledge domains has become pervasive in modern sciences. The interactive dialogs invoked among researchers from different fields have inspired new knowledge and even new fields (Salter & Hearn, 1996; Palmer, 2001; Derry, Schunn, & Gernsbacher, 2005; Lee et al., 2008). Public and private funding agencies have increasingly encouraged interdisciplinary collaborations that involve the integration of knowledge from multiple domains. Interdisciplinary scientific collaboration has also gained much attention in cognitive science (Derry, Schunn, & Gernsbacher, 2005), library and information science (Huang & Chang, 2011), social sciences (Moody, 2004), and health sciences (Lee et al., 2008). These studies have crucial implications in revealing the mechanism of interdisciplinary collaboration and in fostering such collaboration at the level of scientific policies. To study interdisciplinary collaboration, two issues need to be addressed: interdisciplinarity and collaboration.

To explore interdisciplinarity, it's natural to start with the research area or expertise of each author in collaboration. However, this is not a trivial task, especially when examining a massive set of co-authored papers. Previous studies utilized manual or semi-manual labeling to approximate the area or expertise of each author (Chua & Yang, 2008). Chua and Yang (2008) identified authors' areas based on their department, division, center, or institution noted in the addresses. Compared to purely manual labeling, this method provides a relatively consistent and reproducible way of identifying a proxy for authors' areas

or expertise, but the information in the addresses can only give a general idea of these details. In this paper, topic modeling that emerged in the field of natural language processing is adopted to approximate the authors' areas or expertise. The proposed topic model in this paper is certainly not a perfect solution, as it cannot pinpoint exactly each author's area or expertise. However, it is an improvement over previous solutions and gives a better proxy. For the purpose of brevity, we refer to the proxy of authors' areas or expertise as "expertise" in the rest of the paper. Topic modeling has been widely used to extract latent topics from literatures (Hofmann, 1999; Blei, Ng, & Jordan, 2003). In this paper, we integrate the authors' expertise into the coauthorship network. The authors' expertise can be extracted from their papers using the topic model proposed in the work of Tang, Jin, and Zhang (2008). The extracted topics used to annotate the expertise of authors in the global collaboration network are then incorporated into our two proposed algorithms to generate diversity subgraphs.

For capturing collaboration, the coauthorship network built from multi-authored publications is a widely used proxy (Newman, 2001a; Newman, 2001b). Previous studies have generally focused on investigating the global topology of the coauthorship networks at the macro level (Barabási et al., 2002; Leydesdorff & Wagner, 2008; Moody, 2004; Newman, 2001a, 2001b) or on ranking the influence of individual authors in coauthorship networks at the micro level (Liu, Bollen, Nelson, & Van de Sompel, 2005; Yan & Ding, 2009; Yan, Ding, & Zhu, 2010). However, most of these studies have not explored collaborative relationships between two coauthors from the meso level by analyzing the subgraphs between two specific authors. The meso view of a collaboration network is a conceptual construction with subgraphs as the technical representation. While a micro view takes a single edge (representing coauthorship) or node (representing a coauthor) as the unit of analysis in a coauthorship network, a macro view computes topological metrics for the collaboration network as a whole. Neither the micro nor macro views can effectively capture a local and contextual view of how two authors collaborate. For example, in the micro view, a single weighted edge in a collaboration network may indicate that the two authors connected by the edge have co-authored five papers, while totally neglecting the fact that another coauthor was involved in all five papers. In the macro view, we may find that when averaged over all the pairs of coauthors in the global network, each pair of coauthors would co-author five papers, while actually one group of authors typically co-author one paper and another group of authors typically co-author 15 papers. In order to address these problems, this study proposes two algorithms to generate a diversity collaborative subgraph based on multiple paths between two coauthors, where the meso perspective on scientific collaboration is explored by looking into local collaborative context between two given authors.

Based on the expertise-annotated coauthorship network, we define the diversity subgraph and propose the Diversity Subgraph Extraction (DSE) algorithm for extracting the diversity subgraph between two specific authors from the global coauthorship network. The diversity subgraph gives a meso view of the interdisciplinary collaborative relationship between two authors by maximizing the amount of different expertise covered under the constraint on the size of the subgraph. We also propose an adapted algorithm, Constraint-based DES (CDSE), which can further narrow the meso perspective diversity subgraph to only include the paths with predefined intermediate authors. These two algorithms can extract informative cross-domain collaboration subgraphs between two authors from co-authorship graphs with hundreds or thousands of nodes. These algorithms thus help us to focus on the meaningful collaboration patterns without getting lost in huge graphs.

As seen in Figure 1, in a coauthorship network with each author labeled with his or her expertise, "John" and "Mary" are connected through a large number of co-authors who specialize in a number of fields (Figure 1a). Figure 1b is constructed to include the selection and presentation of influential and representative intermediate authors. We can see that the identified diversity subgraph between John and Mary (Figure 1b) presents a meso view of the multi-disciplinary relationship between John and Mary. Compared to Figure 1a, the diversity subgraph in Figure 1b is more appropriate for visualization.

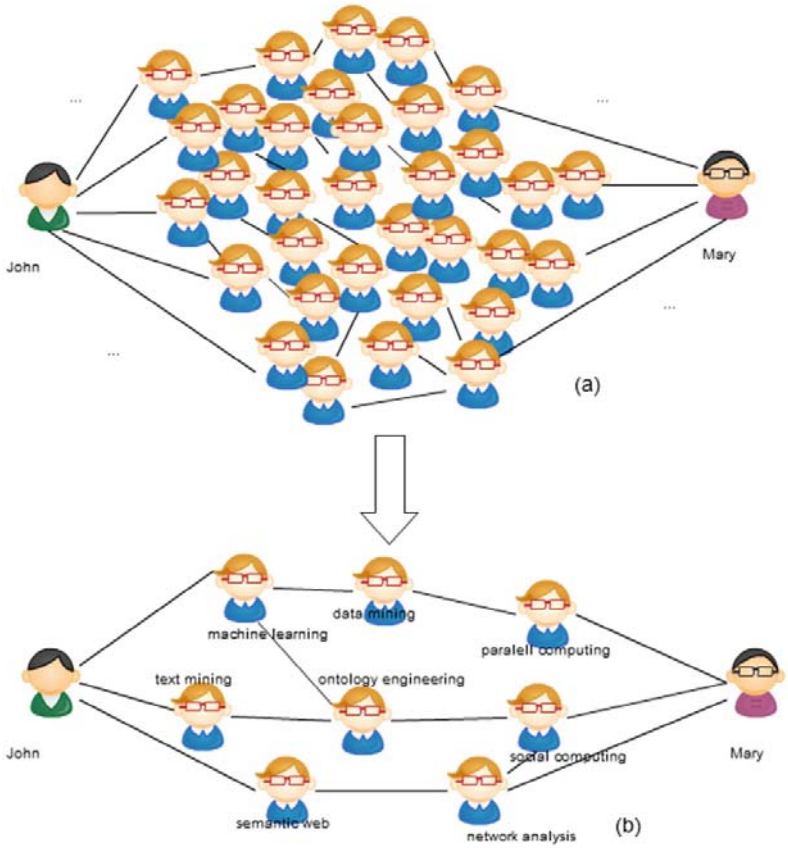

Figure 1. The diversity subgraph forms the set of top-k thematically-diverse paths between John and Mary.

2 Related work

Scientific collaboration

Due to advances in data sources, computing facilities, and software, coauthorship analysis can now target large-scale networks. Researchers have investigated these networks' graph size, largest components, geodesic distance, degree distribution, clustering coefficient, centrality, and k-core, as well as their dynamics over time. Barabási et al. (2002) studied the evolution of the coauthorship network in mathematics and neuroscience over an eight-year period (1991-1998) using its size and structural characteristics, and built a model to simulate the structural mechanisms that govern its evolution. Moody (2004) took advantage of variations in the global topology of the coauthorship network in sociology to reveal the field's research practices in the last 30 years. Meanwhile, another class of studies has developed different indicators of the influence of authors/institutions/countries through analyzing coauthorship network properties from a micro perspective (Börner et al., 2005; Liu et al., 2005; Yan &

Ding, 2009) using centrality measures to analyze different levels of integration (e.g. authors, institutions, countries). However, few studies have provided a meso perspective on scientific collaboration networks. Given an individual paper, Shi et al. (2010) constructed a subgraph of citation network for papers in the reference list of the given paper and investigated the pattern of citing behaviors. For two specific coauthors, He et al. (2011) built a contextual subgraph of coauthorships between these two authors and analyzed the coauthoring patterns for prestigious authors. However, neither of the two studies has incorporated authors' expertise into the investigation.

Subgraph detection

A group of related work has focused on subgraph detection. Faloutsos et al. (2004) defined the Connection Subgraph Problem as follows: given an edge-weighted undirected graph G, vertices s and e from G and an integer budget b, find a connected subgraph H containing s and e and at most b other vertices that maximize a "goodness" function $g(H)$. They also proposed an electrical circuit-based analogy as the goodness function. Ramakrishnan (2005) adapted Faloutsos' algorithm (2004) with heuristics for edge weighting that depends indirectly on the semantics of the entity and property types in the ontology and on characteristics of the instance data. These studies have provided useful techniques for detecting semantic associations, but none of previous studies have addressed the problem of detecting diverse subgraphs in collaboration networks annotated with authors' expertise.

Topic modeling

Since the introduction of Latent Dirichlet Allocation (LDA) presented by Blei, Ng, and Jordan in 2003, various extended LDA models have been proposed by researchers in different domains, i the Author-Topic model (AT) (Rozen-Zvi, Griffiths, Steyvers, & Smyth, 2004). In addition to extracting the subject content of the documents, the AT model can depict the research interests or expertise of authors simultaneously. Tang et al. (2008) proposed an extended model based on the AT model, namely, the Author-Conference-Topic model (ACT), which calculates the probability of a topic for a given author, the probability of a word for a given topic, and the probability of a conference for a given topic. Some studies have applied topic models to graph data mining, including community detection and evaluation (Li et al., 2010; Ding, 2011).

Result diversification

Another important group of related work addresses result diversification in document retrieval. Carbonell and Goldstein (1998) used a linear combination of relevance and diversity as the objective function, while Zhang and Hurley (2008) studied it as a combinatorial optimization problem. Variations of the method used in Zhang and Hurley (2008) have placed a threshold on either relevance or diversity by maximizing one of the two. Although these studies target document search, their methods can be adapted to diversification in path finding.

3 Methods

Since the path-finding problem itself is a NP-hard problem, post-processing of the full set of results is unreasonable in large-scale networks. Thus heuristics-based optimization of the diversity function is designed herein to perform simultaneously with the process of path finding. An electronic circuit analogy is used to measure the strength of paths (Faloutsos et al., 2004), which can theoretically be viewed as an adapted version of random walk. Intuitively, this method values the path that most easily carries extensive information flow.

3.1 Topic modeling

In order to annotate the coauthorship network, the ACT model proposed by Tang et al. (2008) is applied to capture authors' expertise. Intuitively, the ACT models the following process: authors decide to write a paper with certain topics; according to those topics, a set of words are used to describe the study, and an appropriate venue is selected for publishing this paper. The input of the model is the title, abstract, or full text of papers, authors, and publication venues (e.g. journals, conference). The output contains probability distributions of author-over-topic, paper-over-topic, and conference-over-topic. Mathematically, the ACT model is a hierarchical Bayesian network. Fixed hyperparameters α, β, and μ (α=50/T, β=0.01, and μ=0.1) characterize people's uncertainty about the parameters (i.e., θ, ϕ, and ψ) of prior Dirichlet distributions. Using Bayes' rule and multinomial conditional distributions, the posterior distributions are calculated and further estimated by Gibbs sampling. The Gibbs sampling method is used to generate random samples from a joint distribution. It is especially applicable in situations where the joint distribution is not fully known or hard to deal with directly, while being easy to sample from the conditional distributions of a random variables subset. Gibbs sampling iteratively generates a sample of one variable or a subset of random variables from their conditional distribution on the current values of other variables in the joint distribution. Although those samples are dependent, it can be shown mathematically that the limit distribution of those samples is the same as the joint distribution that we target (Lawrence et al., 1993; Liu, 1994).

The probability of a word given a topic ϕ, the probability of a conference given a topic ψ, and the probability of a topic given an author θ can be estimated as:

$$\phi_{zw_{di}} = \frac{n_{zw_{di}} + \beta_{w_{di}}}{\sum_{w_v}(n_{zw_v} + \beta_{w_v})}$$

$$\psi_{zc_d} = \frac{n_{zc_d} + \mu_{c_d}}{\sum_{c}(n_{zc} + \beta_{c})}$$

$$\theta_{xz} = \frac{m_{xz} + \alpha_{z}}{\sum_{z'}(m_{xz'} + \alpha_{z'})}$$

in which *d* represents documents, *w* stands for words, *x* for author, *z* for topic, and *c* for publication venue.

3.2 Construction of diversity subgraph

In this section, the problems in finding the (constraint-based) subgraph formed by the set of top-k thematically diverse paths are formalized. We define the following two problems: (1) diversity optimization for connected subgraphs, and (2) diversity optimization for constraint-based connected subgraphs.

3.2.1 Diversity Optimization for Connected Subgraph Problem

Given: an edge-weighted undirected graph *G* with nodes labeled with classes, source *s,* and sink *e* from *G*.

Find: a connected subgraph G_s composed of the top-k paths between s and e that maximizes the diversity function $D(G_s, k)$.

More paths generally add to the diversity of the set. Thus the diversity function $D(G_s, k)$ implies a competing balance between relevance and diversity, using the minimum number of paths (i.e. most informative paths) to cover the maximum number of different topics.

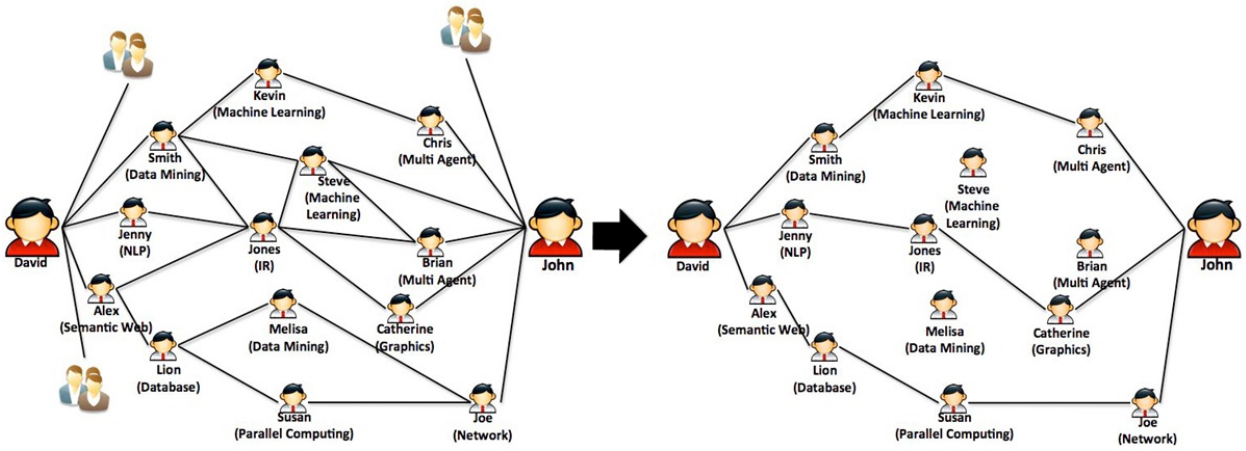

Figure 2. Example of diversity optimization for connected subgraph problem.

As shown in Figure 2, in a scientific collaboration network, we want to find the diversity subgraph between "David" and "John." In this figure, we generate a subgraph with the top three paths between David and John according to the number of co-authors and diverse research areas. Other authors, such as "Steve," "Brian," and "Melisa." are excluded because they have more distant or weaker collaborative relationships with authors in the subgraph, and their expertise is already covered by existing authors in the diversity subgraph.

The DSE algorithm includes three major steps:

- *The pre-processing step (or candidate graph generation):* in candidate graph generation, a smaller graph containing the source and sink vertices is created based on heuristics (e.g. node degree). Nodes in the general neighborhood of the source and the sink and with higher heuristic value are favored in the candidate graph. The candidate graph captures as many relevant nodes as possible, while restricting computing costs to an acceptable level;

- *Electrical network computation:* the candidate graph is viewed as an electrical circuit. According to Ohm's law and the conservation of electricity, the voltage of each node and the current of each edge are obtained by solving a system of linear equations. Currents carried by all the source-to-sink paths are calculated (Faloutsos et al., 2004); and

- *Diversity subgraph generation:* paths that carry the larger amount of current and have more new nodes are selected in an iterative process. In each iteration, the path that scores the highest marginal current per number of existing types of nodes is selected and added to the diversity subgraphs.

For candidate graph generation, the algorithm first adds the source and sink vertex of the candidate graph, then creates a buffer vertex set B, which stores all the adjacent vertices of nodes in the candidate graph C. In each iteration, the vertex with the highest degree is selected and added to the candidate graph, while all its adjacent vertices are added to the buffer vertex set. Once the number of vertices in the candidate graph reaches a given threshold, edges of those vertices from the original graph are updated.

For electricity calculation, the voltage and current in the candidate graph are based on two basic rules of the electrical circuit:

- Ohm's law: the current through a conductor between two points is directly proportional to the potential difference or voltage across the two points, and inversely proportional to the resistance between them. It can be denoted mathematically as:

$$I = \frac{U}{R}$$

- For any point in the current circuit, the amount of current that goes into the point equals the amount that goes out of the point, such as:

$$\sum I_{in} = \sum I_{out}$$

It is worth noting that the weights of edges are modeled in the electricity circuit as a resistance. In this study, the total amount of current in a subgraph is taken as the measure of the informativeness/importance of the paths in the subgraph with respect to the collaborative relationship between a specific pair of authors. This is a good measure because the current incorporates and quantifies large amount of structural information. First, the current at this level has taken the length of the paths into consideration. Longer paths have larger resistance and are thus less favored. Second, the information flow is also included in the amount of current. Assuming an equal amount of information goes out from one node through each of its paths, the information flow quantifies the amount of information carried by one path. Thus if one path has many out-going branches, the information flow going through it is diluted. This can be naturally characterized by the current measure. Information flow does not consider the length of paths, and path length does not consider branching structures of paths. However, the amount of current takes both into consideration.

For the diversity subgraph generation, the diversity function is defined as follows:

$$D(Gs,k) = \arg\max \sum_{i=0}^{k} C_{up}(p_i) / (\sum_{v \in V_{pi}} \delta_{iv} + 1)$$

where δ_{iv} is a binary coefficient, and δ_{iv} =0 if G_s does not contain nodes with the same type of vertex v, otherwise, δ_{iv}=1. The algorithm maximizes the marginal current per existing node type and takes both the diversity and informativeness of paths into consideration. The diversity function used in this paper balances between the strength of the path (i.e. relevance) and coverage of novel paths (i.e. diversity).

3.2.2 Diversity Optimization for Constraint-based Connected Subgraph Problem

Given: an edge-weighted undirected graph G with nodes labeled with classes, source s, sink e, and constraint vertex c from G.

Find: a connected subgraph G_{sc} composed of the top-k paths between s and e via constraint vertex c that maximizes the diversity function $D(G_{sc}, k)$.

For example, researcher A may want to explore the potential ways of establishing a collaborative relationship with a well-known expert B through a certain person C. This can be formalized as the problem of extracting the diversity subgraph between A and B with the current collaborator C as a constrained intermediate node.

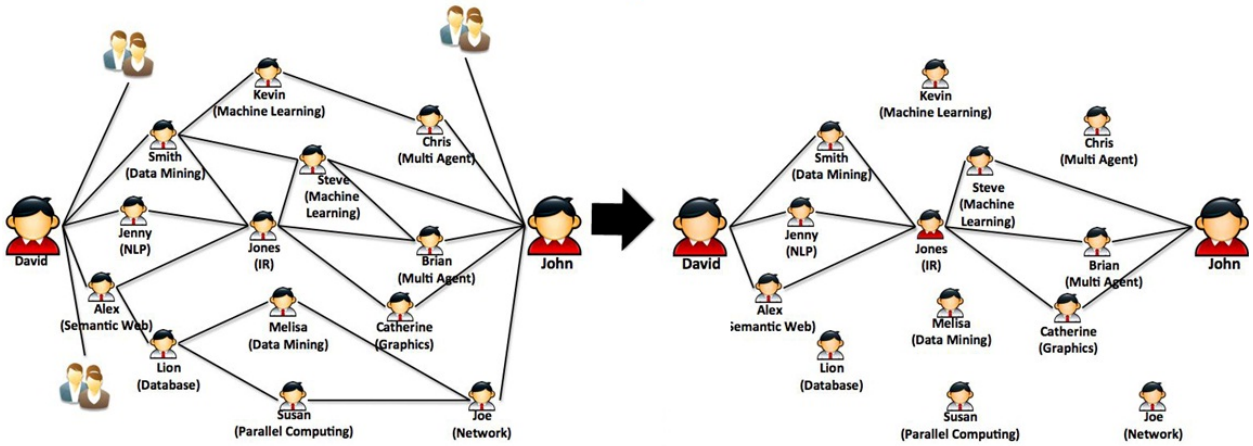

Figure 3. Diversity optimization for constraint-based connected subgraph problem.

In Figure 3, suppose we want to find the subgraph between "David" and "John" via "Jones" with the research area "Information Retrieval." By considering coauthors' preferences (i.e. Jones), a subgraph composed of a set of top-k paths between David and John via Jones is extracted, which maximizes the diversity function. Figure 3 shows that all the paths go through Jones and authors with similar expertise, and weaker collaborations are not included in this extracted diversity subgraph.

In this section, the problem definition and DSE algorithm are further extended to take coauthors' preferences into consideration as a constraint node in order to create a constraint-based diversity subgraph extraction (CDSE) algorithm. A subgraph connecting the source and the constraint node and a subgraph connecting the constraint node and the sink are generated. These two subgraphs are further merged according to a merging mechanism. A node having a higher degree intermediates on more paths with a high current flow and hence needs to be retained. This is how the merging mechanism works: for all shared nodes, the node with a higher degree is retained in that particular subgraph and is discarded in the other subgraph. All duplicate edges are also removed.

After merging the two subgraphs, we generate a new graph with the number of paths between k and k^2. This merged graph is seen as a new candidate graph. The voltages and current are recalculated. Using the DSE algorithm, the subgraph G_{sc} with top k paths between source node s and sink node e via the constraint node c is generated.

3.3 Datasets for demonstration

The DSE and CDSE algorithms are implemented in an academic coauthor network extracted from the academic search system ArnetMiner (Tang et al., 2007; Tang et al., 2008) in the computer science field. The coauthor data set consists of 640,134 authors and 1,554,643 coauthor linkages. In the coauthorship network, nodes represent authors and edges represent coauthorship weighted by the number of co-authored papers. The edge weights are taken as the inverse of the resistor in the electrical network computation in the experiments. Additionally, the titles of about 230,000 papers associated with those authors are used as input for the ACT model, which generates a probability distribution over topics for each author and a set of representative words for each topic. Hot topics include natural language processing, Semantic Web, machine learning, support vector machines, and information extraction. For each author, the topic in which he/she has the highest probability is taken as his/her expertise label.

4 Results

4.1 Diversity subgraph Extraction (DSE) Algorithm

For the coauthorship network, two use cases of diversity subgraphs between prestigious researchers are provided. Figure 4 shows the diversity subgraph between Jiawei Han (expert in data mining) and James Hendler (expert in Semantic Web) and Table 1 presents the top-ranked path in Figure 4. Authors in Figure 4 specialize in 19 different subject areas. Figure 4 sketches the various interdisciplinary collaborations between Jiawei Han and James Hendler, including data mining, machine learning, Semantic Web, and ontology engineering.

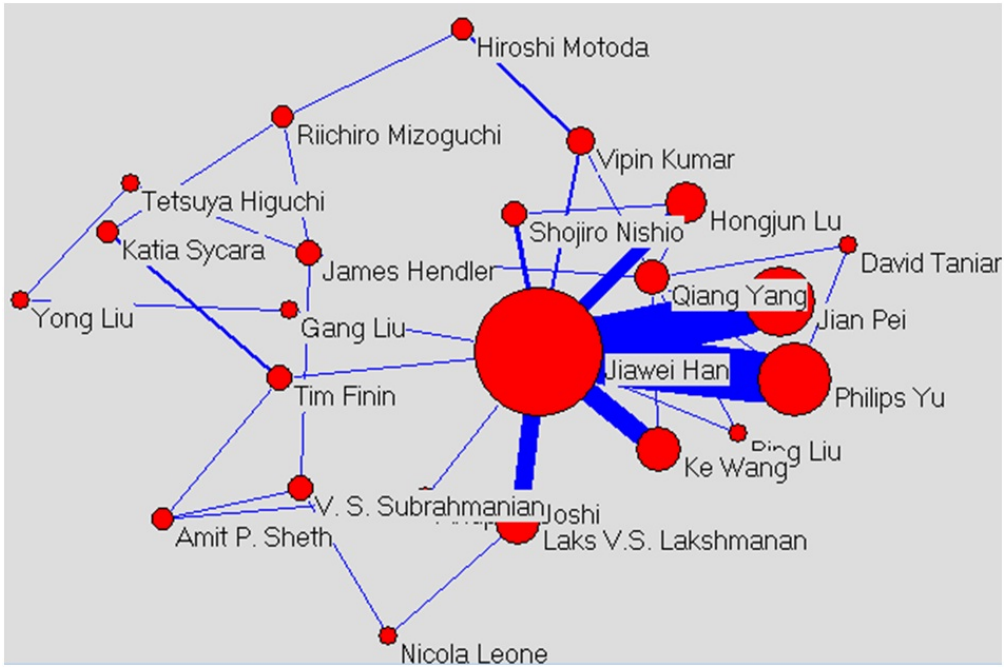

Figure 4 Thematically diversity subgraph between Jiawei Han and James Hendler.

As shown in Table 1, paths containing coauthors with stronger collaborative relationships are ranked higher by current. Shorter paths generally rank higher than longer paths.

Table 1. Ranking of paths in the subgraph between Jiawei Han and James Hendler.

Paths	**Current**
James Hendler→V. S. Subrahmanian→Laks V.S. Lakshmanan→Jiawei Han	0.0042627
James Hendler→V. S. Subrahmanian→Nicola Leone→Laks V.S. Lakshmanan →Jiawei Han	0.0092124
James Hendler→Qiang Yang→David Taniar→Philips Yu→ Jiawei Han	0.0048753
James Hendler→Tetsuya Higuchi→Yong Liu→Gang Liu→ Jiawei Han	0.0043287
James Hendler→Riichiro Mizoguchi→Hiroshi Motoda→Vipin Kumar→Jiawei Han	0.0031634

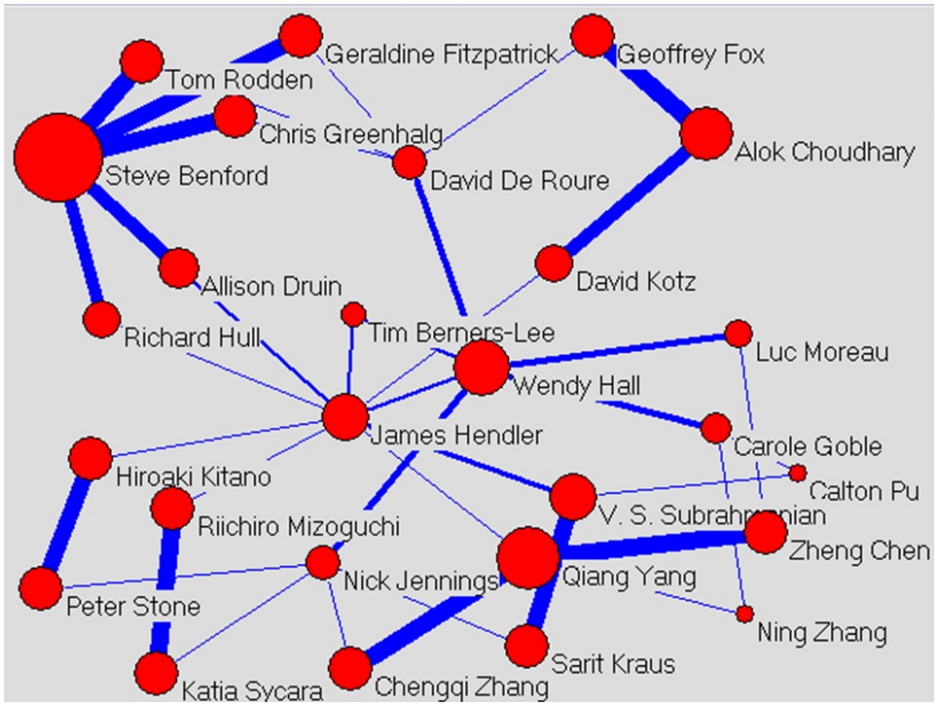

Figure 5. Thematically diversity subgraph between Tim Berners-Lee and James Hendler.

Figure 5 shows the diversity subgraph between Tim Berners-Lee and James Hendler (both experts in Semantic Web and World Wide Web). The authors presented in this subgraph are distributed over 18 different subject areas, including Semantic Web, data mining, information retrieval, and ontology engineering.

Table 2. Ranking of paths in the subgraph between Tim Berners-Lee and James Hendler.

Paths	Current
Tim Berners-Lee →James Hendler	1
Tim Berners-Lee → Wendy Hall →James Hendler	0.611008
Tim Berners-Lee→Wendy Hall→David De Roure →Geraldine Fitzpatrick →Steve Benford→Allison Druin→James Hendler	0.0133855
Tim Berners-Lee→Wendy Hall →David De Roure →Chris Greenhalgh→Steve Benford→Allison Druin →James Hendler	0.0100913
Tim Berners-Lee→Wendy Hall→David De Roure →Tom Rodden→Steve Benford→Allison Druin →James Hendler	0.007547

Table 2 shows the most informative and diverse paths between Tim Berners-Lee and James Hendler. It is reasonable that the direct collaboration between these two authors ranks highest. Notably, Wendy Hall exists in four of the top five diverse and informative paths. This is due to the fact that Tim Berners-Lee and Wendy Hall as well as James Hendler and Wendy Hall both have close and strong collaborations. The large amount of current carried through these edges is so overwhelming that multiple paths with Wendy Hall are continuously added to the diversity subgraph.

4.2 Constraint-based Diversity subgraph Extraction (CDSE) Algorithm

Constraint-based subgraphs advance the DSE algorithm by taking coauthors' preferences into consideration as a constraint node in the path. For the coauthorship network, two use cases of diversity subgraphs between prestigious researchers are provided. Figure 6 shows the constraint-based subgraph between Eugene Charniak (expert in natural language processing) and Steffen Staab (expert in Semantic Web) with Susan Dumais (expert in information retrieval) as the constraint node. This diversity subgraph involves 19 different subject areas involving natural language processing, Semantic Web, and knowledge management (Table 3).

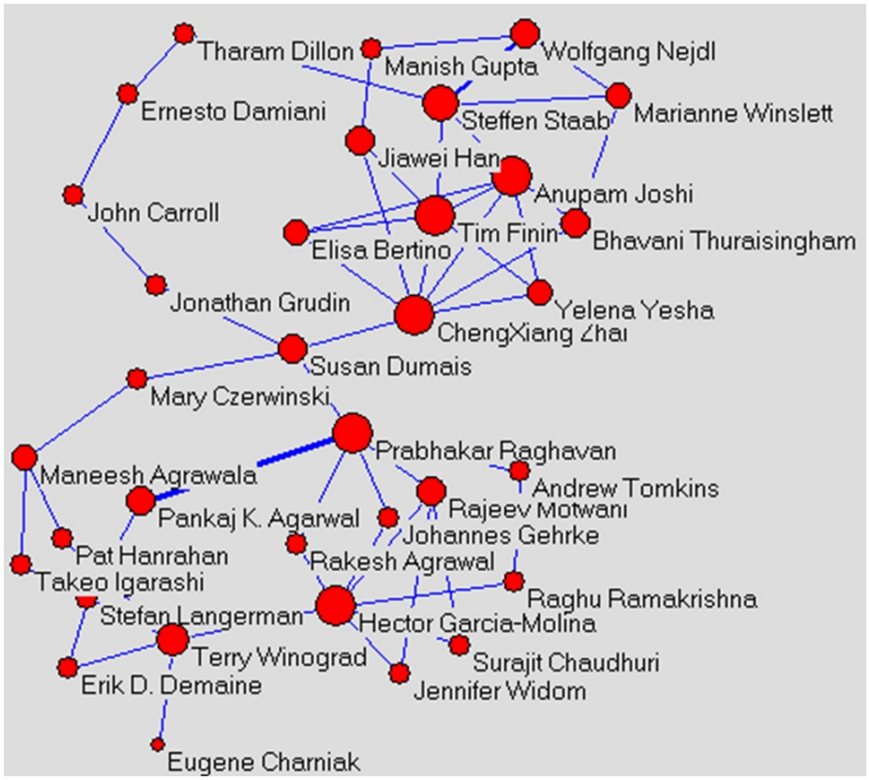

Figure 6. Constraint-based subgraph between Eugene Charniak and Steffen Staab with constraint on Susan Dumais.

Table 3. Ranking of paths in the subgraph between Eugene Charniak and Steffen Staab with constraint on Susan Dumais.

Paths	current
Eugene Charniak→ Terry Winograd → Pat Hanrahan →Maneesh Agrawala →Mary Czerwinski →Susan Dumais→Chengxiang Zhai→Tim Finin→Steffen Staab -	0.0302355
Eugene Charniak →Terry Winograd →Takeo Igarashi →Maneesh Agrawala →Mary Czerwinski→Susan Dumais→Chengxiang Zhai →Anupam Joshi →Steffen Staab	0.0302355
Eugene Charniak →Terry Winograd →Takeo Igarashi →Maneesh Agrawala →Mary Czerwinski →Susan Dumais →Chengxiang Zhai →Tim Finin→Steffen Staab	0.0302355
Eugene Charniak →Terry Winograd → Pat Hanrahan →Maneesh Agrawala →Mary Czerwinski →Susan Dumais →Chengxiang Zhai→Anupam Joshi →Steffen Staab	0.0302355
Eugene Charniak →Terry Winograd →Hector Garcia-Molina →Johannes Gehrke →Prabhakar Raghavan →Susan Dumais →Chengxiang Zhai →Anupam Joshi →Steffen Staab	0.015775

Figure 7 shows the thematically diversity subgraph between Jiawei Han and Tim Berners-Lee with James Hendler as the constraint node. Table 4 presents the top-ranked paths between Jiawei Han and Tim Berners-Lee with James Hendler as the constraint node.

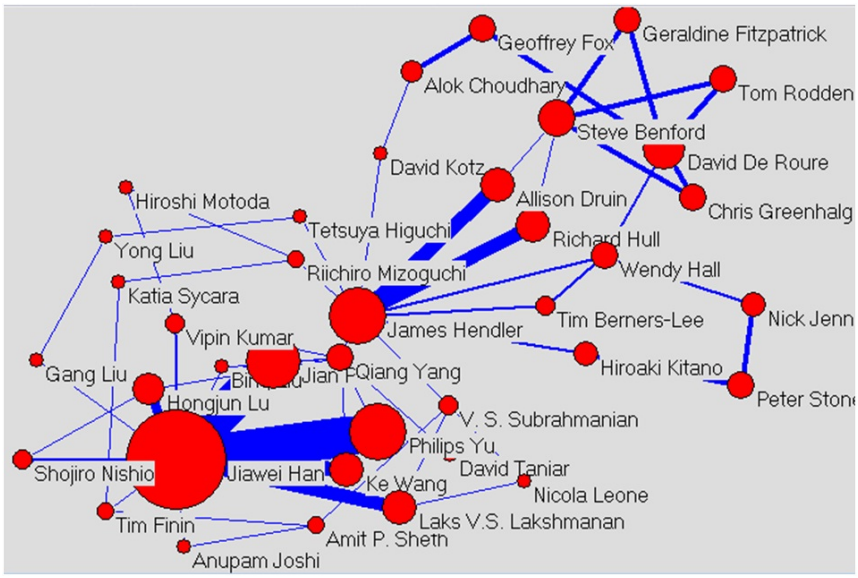

Figure 7. Constraint-based subgraph between Jiawei Han and Tim Berners-Lee with constraint on James Hendler.

Table 4. Ranking of paths in the subgraph between Jiawei Han and Tim Berners-Lee with constraint on James Hendler.

Paths	current
Tim Berners-Lee→James Hendler→V. S. Subrahmanian→Laksv. S. Lakshmanan→Jiawei Han	0.0652928
Tim Berners-Lee→James Hendler→Qiang Yang→Hongjun Lu→Jiawei Han	0.0513285
Tim Berners-Lee→James Hendler→Qiang Yang→Bing Liu→Jiawei Han	0.0427738
Tim Berners-Lee→James Hendler→Qiang Yang→Ke Wang→Jiawei Han	0.0427738
Tim Berners-Lee→James Hendler→QiangY ang→Jian Pei→Jiawei Han	0.0427738

5 Discussion

In this section, the proposed DSE algorithm is evaluated against the BFS-based subgraph extraction to quantify the performance of DSE in capturing the diversity in interdisciplinary collaboration. Additionally, limitations of the proposed framework are discussed.

Evaluation

The DSE algorithm forms the core part of our proposed methods. In this section, the DSE algorithm is evaluated against Breadth First Search (BFS) by comparing the diversity of expertise of the resulting subgraphs. The top 25 authors in our dataset are selected based on their citation counts. There are 300 possible ways of pairing them. For each pair of authors, a diversity subgraph is extracted from the proposed DSE algorithm and a baseline subgraph is retrieved using the BFS. The baseline subgraph for a pair of authors consists of all the paths that connect the pair of authors. In order to quantitatively access the proposed algorithms, a *diversity index*, defined as the fraction of the number of covered topics over the graph size, is used to quantify the diversity level of a subgraph. The following formula gives us a measure of the diversity level of a subgraph S:

$$Diversity(S) = \frac{|S_{type}|}{|S|},$$

$|S|$ is the number of nodes in the subgraph, and $|S_{type}|$ is the number of types covered in the subgraph. The difference index between the diversity levels of two subgraphs S_1 and S_2 is calculated by the following formula:

$$Difference(S_1, S_2) = \frac{Diversity(S_2) - Diversity(S_1)}{\max(Diversity(S_1), Diversity(S_2))}$$

The side-by-side box plots for the diversity index for BFS-based subgraphs, the diversity index for DSE-based subgraphs, as well as the difference index comparing DSE to BFS are shown in Figure 8. As seen in Figure 8, the median diversity index for BFS-based subgraphs is lower than that of DSE-based subgraphs; and the median difference index and the difference indexes for the bulk of cases are above zero (dented line) in comparing DSE to BFS. These indicate that diversity subgraphs cover a relatively larger number of topics based on the relatively smaller size than those covered by baseline subgraphs. We perform a two-sided paired t-test on the diversity indexes for the two types of subgraphs, the diversity subgraph and the baseline subgraph, wherein:

- The null hypothesis is that the mean diversity index in baseline graphs is the same as that of diversity subgraphs; and

- The research hypothesis is that the mean diversity index in baseline graphs is less than that of the diversity subgraphs.

The p value for this test is less than 0.001 and the corresponding 95 percent confidence interval for the estimated difference of average diversity index in DSE and BFS is [0.047, 0.076], indicating that we can reject the null hypothesis and conclude that the diversity index is significantly less than that of the diversity subgraph. A second one sample t test is also conducted to test the null hypothesis that the difference index is equal to 0 versus the alternative hypothesis that the difference index is not equal to 0. The resulting p value is less than 0.001 and the mean difference index is 0.140 with a 95 percent confidence interval [0.107, 0.165], suggesting that the average diversity index increases 14.0 percent comparing DSE to BFS. In our experiment, the proposed diversity subgraph thus carries richer information on the thematic diversity with smaller size than the corresponding BFS subgraph.

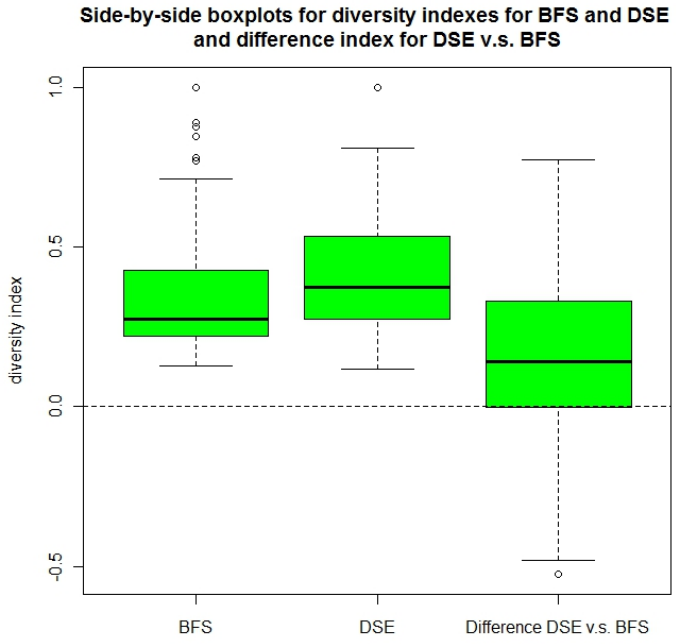

Figure 8. Side-by-side boxplot for diversity indexes of DSE and BFS as well as the difference index between DSE and BFS

Limitations

Limitations of our study include:

- The proposed algorithms contain several approximation procedures, one of which is to use a smaller candidate graph for electric calculation instead of the real global graph. Although we control for bias through generating a customized candidate graph for each inquiry on author pairs, the artificial distortion that these procedures might bring to the results is not yet evaluated;

- A non-trivial task of applying the proposed algorithms to a new dataset is determining how to select or adjust the input parameters, including the measure of relevance in forming a candidate graph, and the size threshold of the candidate graph. We do not provide a detailed guideline herein on how to select those parameters. But simply speaking, the selection of the measure of relevance depends on the question of interest. The size threshold of the candidate graph depends on the tradeoff between the density of the graph and the available computing power; and

- There is no gold standard against which we can quantify and assess the diversity of the diversity subgraphs. Large-scale experiments on datasets from different application areas are thus needed to evaluate the effectiveness of the proposed algorithms.

6 Conclusion

The major contribution of this paper is to propose a novel framework of analyzing interdisciplinary collaboration, which provides an inspection of coauthorship network from an interdisciplinary angle and reveals previously hidden patterns in coauthors' expertise. We define the problem of thematically extracting diversity subgraphs in collaboration networks herein and propose an algorithm for thematically detecting diversity subgraphs between two authors. Furthermore, we extend the algorithm to take into consideration authors' preferences in the form of a constrained intermediate node along the paths between two nodes. The diversity subgraph presents a meso view of the multi-disciplinary collaborative relationships between a pair of coauthors.

This paper integrates authors' expertise obtained from topic modeling into the analysis of scientific collaboration and defines the diversity graph, which presents a meso view of multi-disciplinary collaborative relationships for a pair of coauthors. With advances in modern science, more and more research issues require complementary knowledge and skills from different scientific disciplines. Presenting the expertise associated with authors in collaborative relationships may shed light on specific features of cross-domain collaboration.

We develop algorithms for extracting the diversity subgraph from the global coauthorship network and apply the algorithms in a large coauthorship network in computer science. The algorithms are evaluated against the BFS. Compared to the subgraphs produced by the BFS, the DSE/CDSE algorithms capture more diversity with fewer nodes. Our future work can be expanded to include the thematically diversity subgraphs between more than two nodes and the consideration of more complex constraints that are defined by coauthors. Our approach only investigates a snapshot of the coauthorship network, disregarding rich information on the dynamics of those graphs, which may also lead to a future research topic. Future developments of the proposed methods can be applied in a variety of domains, such as identifying a connection between a drug and a target in cheminformatics or bioinformatics, and locating a meaningful subgraph between two journals or two papers in scholarly communications, given the interests in analyzing the diversity of subgraphs of two nodes.

7 Acknowledgment

This work is supported by NIH-funded VIVO project (NIH Grant U24RR029822). Thanks to Dr. Daifeng Li, Yuyin Sun, and Shanshan Chen for their comments and suggestions. We gratefully acknowledge the helpful suggestions from of anonymous reviewers from the *Journal of Informetrics*.

8 References

Aleman-Meza, B., Halaschek, C., Arpinar, B. I., & Sheth, A. (2003). Context-aware semantic association ranking. In *Semantic Web and Databases Workshop Proceedings*, 33-50, Berlin, Germany.

Alkhateeb, F., Baget, J. F., & Euzenat, J. (2008). Constrained regular expressions in SPARQL. In *SWWS*, 91-99.

Anyanwu, K., Maduko, A., & Sheth, A. (2005). SemRank: ranking complex relationship search results on the semantic web. Paper presented at *the 14th international conference on World Wide Web*.

Anyanwu, K., Maduko, A., & Sheth, A. (2007). Sparq2l: Towards support for subgraph extraction queries in RDF databases. In *Proceedings of the 16th international conference on World Wide Web*, 797-806, New York, NY, USA.

Barabási, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A*, *311*, 590-614.

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

Börner, K., Dall'Asta, L., Ke, W., & Vespignani, A. (2005). Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity, 10*, 57-67.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 335-336, Melbourne, Australia.

Collapudi, S., & Sharma, A. (2009). An axiomatic approach for result diversification. In *Proceedings of the International World Wide Web Conference*, Madrid, Spain, 381-390

Chua A. Y. K., & Yang CC. (2008). The shift towards multi-disciplinarity in information science. *Journal of the American Society for Information Science and Technology, 59*, 2156–2170.

Derry J. S., Schunn, C. D., & Gernsbacher, M. A., (Eds.). (2005). Interdisciplinary collaboration: An emerging cognitive science. New Jersey: Lawrence Erlbaum Associates.

Ding, Y. (2011). Topic-based PageRank on author co-citation networks. *Journal of the American Society for Information Science and Technology, 62*(3), 449-466.

Ding, Y. (2011). Community detection: Topological vs. topical. *Journal of Informetrics, 5*, 498-514.

Drosou, M., & Pitoura, E. (2010). Search result diversification. *SIGMOD Record*, *39* (1), 41-47

Faloutsos, C., McCurley, K. S., & Tomkins, A. (2004). Fast discovery of connection subgraphs. Paper presented at *the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, USA, 118-127

He, B., Ding, Y., & Ni, C. (2011). Mining enriched contextual information of scientific collaboration: A meso perspective. *Journal of the American Society for Information Science and Technology, 62*(5), 831-845.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50-57, Berkeley, CA, USA.

Huang, M., & Chang, Y. (2011). A study of interdisciplinarity in information science: Using direct citation and co-authorship analysis. *Journal of Information Science, 37*(4), 369-378.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., & Wootton, J. C. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, *262,* 208-214.

Lee, E. S., McDonald, D. W., Anderson, N., & Tarczy-Hornoch, P. (2008). Incorporating collaborator concepts into informatics in support of translational interdisciplinary biomedical research. *International Journal of Medical Informatics, 78(*1), 10-21.

Leydesdorff, L., & Wagner, C. S. (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics, 2*(4), 317-325.

Li, D., He, B., Ding, Y., Tang, J., Sugimoto, C., Qin, Z., et al. (2010). Community-based topic modeling for social tagging. Paper presented at *the Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, Toronto, Ontario, Canada.

Liu, X. M., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2005). Coauthorship networks in the digital library research community. *Information Processing & Management, 41*(6), 1462-1480.

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association, 89*, 958-966.

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

Newman, M. E. J. (2001a). Scientific collaboration networks: I. Network construction and fundamental results. *Physical Review E, 64*(1), 016131.

Newman, M. E. J. (2001b). Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality. *Physical Review E, 64*(1), 016132.

Palmer, C.L. (2001). *Work at the boundaries of science: Information and the interdisciplinary research process.* Dordrecht: Wolters Kluwer.

Ramakrishnan, C., Milnor, W. H., Perry, M., & Sheth, A. (2005). Discovering informative connection subgraphs in multi-relational graphs. *SIGKDD Exploration Newsletter 7*(2), 56-63.

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*, Arlington, VA, USA, 487-494.

Salter, L., &, Hearn, A. (1996). *Outside the lines*. Montreal, Canada: McGill-Queen's University.

Shi, X., Leskovec, J, & McFarland, D. A. (2010). Citing for high impact. *In Proceedings of the Joint Conference of Digital Library*, Gold Coast, Australia, 49-58

Tang, J., Jin, R., & Zhang J. (2008). A topic modeling approach and its integration into the random walk framework for academic search. In *Proceedings of 2008 IEEE International Conference on Data Mining (ICDM2008)*, 1055-1060, Pisa, Italy.

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,* 990-998, Las Vegas, NV, USA.

Tang, J., Zhang, J., Zhang, D., Yao, L., Zhu, C., & Li, J. (2007). ArnetMiner: An expertise oriented search system for Web community. *Semantic Web Challenge.*

Yan, E., Ding, Y., & Zhu, Q. (2010). Mapping library and information science in China: A coauthorship network analysis. *Scientometrics, 83*(1), 115-131.

Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology, 60*(10), 2107-2118.

Yan, E., Ding, Y., Milojevic, S., & Sugimoto, C. R. (2012). Topics in dynamic research communities: An exploratory study for the field of information retrieval. *Journal of Informetrics*, *6*(1): 140-153

Zhang, M., & Hurley, N. (2008). Avoiding monotony: Improving the diversity of recommendation lists. In *Proceedings of the 2008 ACM conference on recommender systems,* 123-130, Lausanne, Switzerland