Overlaying communities and topics: An analysis on publication networks

Erjia Yan, Ying Ding, and Elin K. Jacob

{eyan, dingying, ejacob}@indiana.edu
School of Library and Information Science, Indiana University, Bloomington, USA

Abstract

Two layers of enriched information are constructed for communities: a paper-to-paper network based on shared author relations and a paper-to-paper network based on shared word relations. K-means and VOSviewer, a modularity-based clustering technique, are used to identify publication clusters in the two networks. Results show that a few research topics such as webometrics, bibliometric laws, and language processing, form their own research community; while other research topics contain different research communities, which may be caused by physical distance.

Introduction

Research topics such as community detection and topic identification are becoming appealing in bibliometrics. Essentially, these two topics are not new to bibliometricians. On one hand, methods such as hierarchical clustering and k-means have been used to group actors (journals, authors, etc.) in scholarly networks. On the other hand, techniques such as author co-citation analysis (ACA) have been used to identify disciplines or research specialities. These classic tools, however, have obvious limitations, for example, choosing the number of clusters can be arbitrary for hierarchical clustering and k-means; ACA is confronted with several technical limitations (White & McCain, 1998) and issues related to author order selections (first author only, last author only, all authors, etc.) further shadows this technique.

Built upon previous endeavours, the current community detection methods use distinguishable measurements, such as modularity (Newman & Girvan, 2004) and conductance (Leskovec, Lang, Dasgupta, & Mahoney, 2008) to measure the quality of clustering results, and thus no prior knowledge is required to choose the number of clusters. Qualitatively, a community is a subset of nodes densely connected internally and loosely connected externally. Radicchi et al. (2004) gave a quantitative definition of a community: in a strong community each node has more connections within the community than with the rest of the graph ($k_i^{in}(V) > k_i^{out}(V), \forall i \in V$, where k_i is the degree of node i, V is a subgraph). Another thread of effort has successfully applied topic models to discover topics from text. A topic represents an underlying semantic theme and can be informally defined as an organization of words and can be formally defined as a probability distribution over terms in a vocabulary (Blei, 2007). The two methods are the preludes of ongoing investigations on community detection and knowledge discovery.

Intuitively, knowledge discovery can further be extended to research communities; however, either community detection or topic model alone cannot achieve it. In studies of scholarly communications, community detection methods are usually applied to coauthorship networks where all nodes are homogenous authors, and therefore topics cannot solely be identified via authors' collaboration information; meanwhile, topic models are usually implemented to a large corpus and the outcomes are probability distributions of words to each topic; advanced topic models such as Author-Conference-Topic model (Tang, Zhang, Yao, Li, Zhang, & Su, 2008) can generate an author probability distribution to each topic, but authors belonging to each topic may not necessarily belong to the same community.

In a primary attempt to discover topics for research communities, Li et al. (2010) found that communities and topics are interweaving and co-evolving: that is, a research community can carry several topics, and a topic can consist of different collaboration groups. Therefore, discovering knowledge at community level requires overlaying a topic layer on the scholarly network, and such an approach provides an opportunity to study how topics interact with communities.

Different from Li et al.'s (2010) approach, a novel paper-to-paper network is proposed to overlay communities and topics. The advantage of this network is that it allows the embedding of two relations: one is the shared author relation and the other is the shared title word relation; thus only one community detection method is needed and meanwhile the two clustering results are comparable.

In addition, two methods, k-means and the VOSviewer clustering technique are applied to overlay communities and topics in library and information science (LIS). The former is a representative method for traditional graph partitioning and the latter is a novel technique for modularity-based clustering. The present study, therefore, intends to explore three questions:

- Topic-wise, could topics be derived from research communities;
- Community-wise, are research communities driven by topics; and
- Method-wise, would the modularity based clustering method outperform the traditional graph partitioning method?

The rest sections are organized as follows. The second section introduces related work on modularity-based community detection studies; the third section discusses the data and networks; the fourth section introduces the two methods used in the study; the fifth section shows the clustering results and discusses the topics discovered; the sixth section ends the paper with conclusions.

Related work

Topic identification has become a hot research topic in recent years. Methods such as Latent Dirichlet Allocation (LDA) and its related algorithms have been successfully used in various knowledge discovery tasks. More than a decade ago, scholars in library and information science attempted to address this issue from a different perspective, mainly using the method of co-citation analysis. In an important literature of ACA, for example, White and McCain (1998) identified 12 research topics in information science between 1972 and 1995, where the two biggest specialties are experimental retrieval and citation analysis. Later on, White (2003) proposed Pathfinder networks (PFNETs), and found (PFNETs) outperformed ACA in its ability to produce more readable and interpretable results. Persson (1994) analyzed co-cited authors in the *Journal of the American Society for Information Science* (JASIS) publications from 1986 to 1990 and found the intellectual base of information science has two main branches, bibliometrics and information retrieval. Through ACA, Åström (2010) found that there is an evident distinction between the topics studied by information science and library science authors in that the library science authors are separated from information science authors in the co-citation visualization map, a similar discovery has also been made by Waltman, Yan, and Van Eck (2011) where the authors found information science, library science and scientometrics journals have quite different performance scores measured by a recursive bibliometric indicator. In a review article by Morris and Van der Veer Martens (2007), the authors reviewed different approaches to study research specialties from sociological, bibliographical, communicative, and cognitive perspectives. The research topics and specialities found in above articles will be matched with the findings from the present study.

Previous investigations on graph partitioning are confronted with the difficulty of choosing the number of clusters: it has to be pre-assigned or arbitrarily decided. Modern community detection methods use measurements such as modularity (Newman & Girvan, 2004) and conductance (Leskovec, Lang, Dasgupta, & Mahoney, 2008) to measure the quality of clusters obtained. Effectively divided communities usually have high modularity values. They are densely connected internally between the nodes within modules but loosely connected externally between different modules. Conductance uses a similar definition where it can be described as the ratio between the number of edges inside the cluster and the number of edge leaving the cluster (Leskovec, Lang, & Mahoney, 2010). Well defined communities usually have low conductance values. Richardson, Mucha and Porter (2009) formulated a spectral graph-partitioning algorithm and extended Newman's (2006) bipartitioning methods to tripartitioning methods that allowed two-way and three-way divisions at each recursive step. They found that their method yielded higher-modularity partitions. Farkas, Ábel, Palla and Vicsek (2007) proposed the Clique Percolation Method with weights (CPMw) for weighted networks. The advantage of CPMw is that it allows the overlap of one node into more than one cluster. Donetti and Munoz (2004) introduced a method for community detection that exploited the graph Laplacian matrix combined with hierarchical-clustering techniques. They tested the method on the Zachary karate club network and a coauthorship network comprised of authors at ArXiv.org. They found that their method could maximize the modularity of the output while reducing computational time.

In a broad sense, topic models are special forms of community detection. Topic models follow the principle that the more common words the two entities share, the more similar these two entities are, and thus they can be referred to as the topic-based community detection (Ding, 2011 submitted) which differ from the topology-based community detection methods mentioned above. The basic idea of topic-based community detection is to use latent topics to capture semantic dependencies in the textual information. One well-known topic model is the Probabilistic Latent Semantic Indexing (pLSI) model proposed by Hofmann (1999). Built on pLSI, Blei et al. (2003) introduced a three-level Bayesian network, called Latent Dirichlet Allocation (LDA). Probabilistic models have also been extended to include authorship information. Steyvers et al. (2004) proposed an unsupervised learning technique for extracting both the topics and authors of documents. In their Author-Topic model, authors are modeled as probability distributions over topics. Author-Conference-Topic (ACT) Model, proposed by Tang et al. (2008), further extended Author-Topic model to include conference/journal information. The ACT model utilizes probabilistic models to model documents' contents, authors' interests, and also conference/journal simultaneously.

The above algorithms on topology-based and topic-based community detection can effectively partition actors into clusters or assign words into topics. However, applying either method to one homogenous network alone is not able to identify topics at community level. To address this problem, there is a need to either apply both topology-based and topic-based community detection methods to one network or apply one method to two networks. Li et al. (2010) took the first approach by combining LDA with the Girvan-Newman's community detection algorithm and tested their method on a social tagging data set. In this study, the second approach is chosen in that a topology-based community detection method is applied to two paper-to-paper networks, and questions on the interaction between community and topics can therefore be answered.

Data description

Sixteen representative journals in library and information science (LIS) were selected[1]. These journals were selected based on perception (Nisonger & Davis, 2005) and citation-based

[1] See data and interactive visualizations at http://info.slis.indiana.edu/~eyan/papers/cluster/

rankings. Additionally, only those journals indexed by Thomson Reuters' Web of Knowledge (WoK)[2] were included. Using WoK, all articles published in the selected journals between 1988 and 2007 were then identified, and the results were refined by specifying document type as "article" or "review article". In total, 10,344 articles and 10,579 authors were identified. The size of the largest component (LC) in the coauthorship network was 2,197 and these authors in LC were used as the data set to study publication clusters. Disciplinarity is an important factor affecting the size of the LC. In the four coauthorship networks studied by Newman (2001), Medline has the largest component, with 92.6% of all the authors. Social science disciplines tend to have smaller LCs (Yan, Ding, & Zhu, 2010).

The procedure used was as follows: Search all publications by the 2,197 authors; construct two adjacency matrices (i.e., a paper-author matrix and a paper-word matrix that used words from article titles after removing stop words); multiply paper-author and paper-author transpose as well as paper-word and paper-word transpose to obtain two paper-to-paper matrices; run clustering algorithms on both matrices; and finally match the results. In the paper-to-paper (author) matrix (PPAM), a cell value denotes the number of shared authors; in the paper-to-paper (word) matrix (PPWM), a cell value denotes the number of shared title words. The limitation of this process is that the author names were unable to be disambiguated; words' synonyms were not considered. Figure 1 illustrates matrix formation and network construction, and Table 1 provides details of matrix dimensions.

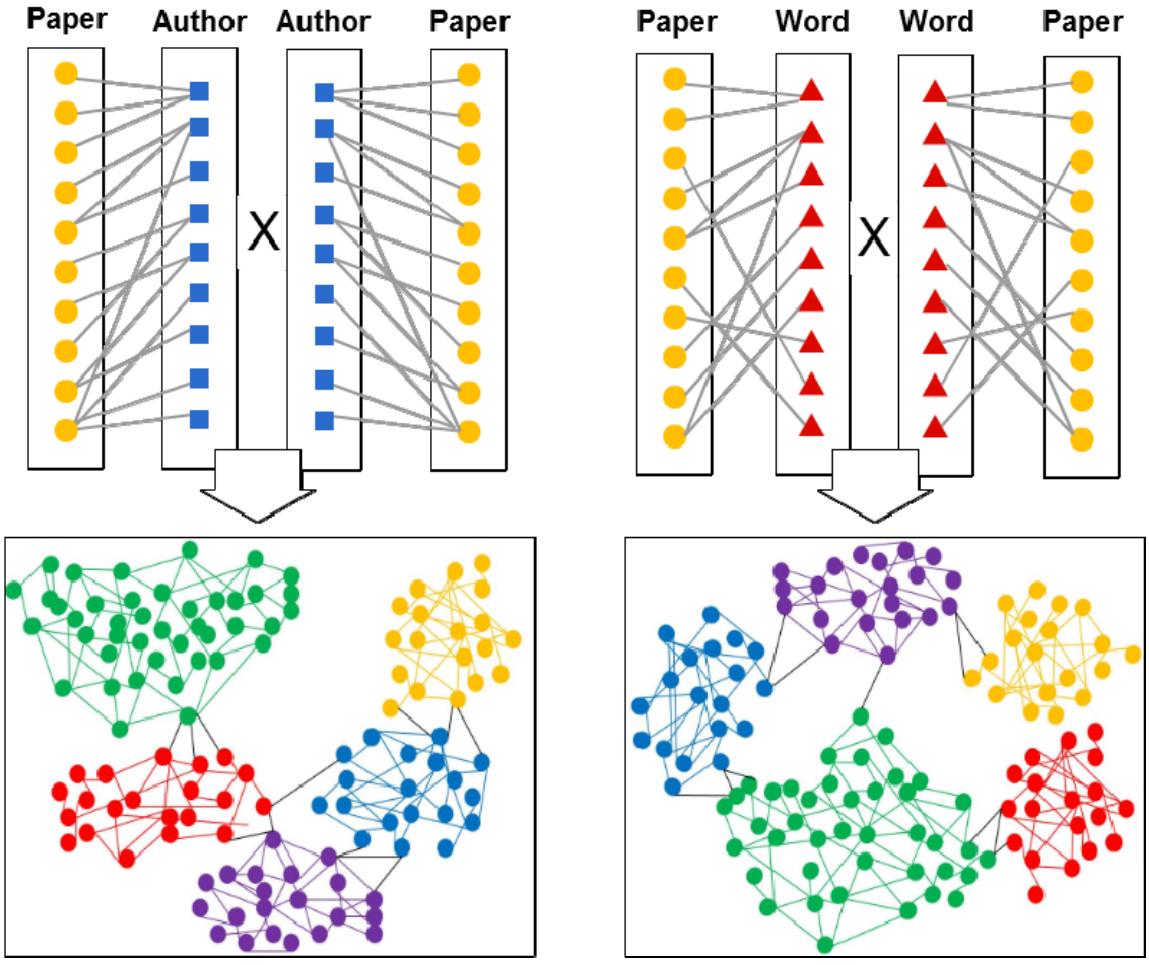

Figure 1. Network construction

Table 1. Network size

[2] http://www.isiwebofknowledge.com/

Number of authors in LC	2,197
Number of articles by authors in LC	3,053
Size of paper-author matrix	3053*2197
Size of paper-word matrix	3053*4449
Size of paper-to-paper matrix	3053*3053

Methods

Fortunato (2010) outlined two phases of research in graph clustering: traditional methods and modularity-based methods. Traditional methods include graph partitioning (e.g., Kernighan-Lin algorithm), hierarchical clustering, partitional clustering (e.g., k-means), and spectral clustering (e.g., algorithms utilizing Laplacian matrices). Modularity-based methods include clustering algorithms that use modules to measure the strength of communities. In this study, k-means was selected to represent traditional clustering methods and VOSviewer[3] was selected to represent modularity-based clustering methods.

Traditional clustering method (k-means)

The cost function of k-means can be denoted as:

$$\sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - c_i\|^2$$

where S_i is the subset of points of the i-th cluster and c_i is its centroid. Each centroid is the mean of the points in that cluster, and the method used to choose the initial cluster centroid positions is to select k observations from X (the data matrix) at random. K-means uses a two-phase iterative algorithm to minimize the sum of point-to-centroid Euclidean distances summed over all k clusters, a.k.a. the cost function.

Traditional clustering methods have several limitations. For example, because hierarchical clustering tends to separate single peripheral vertices from the communities, additional information is needed to understand the real structure of the hierarchies. As Radicchi, Castellano, Cecconi, Loreto and Parisi (2004) have observed, "without such information it is not clear at all whether the identification of a community is reliable" (p. 2659). Graph partitioning and partitional clustering have the limitation that the number of clusters must be specified before implementation. In addition, the treatment of overlapped nodes can be artificial for some graphs (Fortunato, 2010).

Modularity-based clustering method (VOSviewer clustering technique)

VOSviewer clustering technique is selected to represent modularity-based clustering methods. It is developed based on Clauset, Newman, and Moore's (2004) algorithm for weighted networks. Initially, Girvan and Newman (2002) proposed an algorithm that uses edge betweenness to identify the boundaries of communities. An edge with high betweenness is the bridge that interconnects different clusters. The Girvan and Newman (2002) algorithm involves iterative application of four steps: (1) calculates edge betweenness for all edges in the network, (2) removes the edges with highest betweenness, (3) recalculates betweenness for all edges affected by the removal, and then (4) repeats from step 2 until no edge remains. This algorithm is computational time demanding and is optimized into a more efficient algorithm (Clauset, Newman, & Moore, 2004). The new algorithm also incorporated *modularity*, a measurement proposed by Newman and Girvan (2004), to evaluate the

[3] http://www.vosviewer.com/

community structures. The modularity for unweighted network can be calculated as (Newman & Girvan, 2004):

$$Q = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\, e - \|e^2\|$$

In a division of a network into k communities, e_{ij} is the fraction of all edges in the network that link vertices in community i to vertices in community j in the $k*k$ symmetric matrix. The row sums $a_i=\sum e_{ij}$. This quantity measures the fraction of within-community edges minus the expected value of the same quantity in a network with the same community divisions but randomly connected. The aim in community detection is to find the community structure under the maximum modularity Q.

In weighted networks, each cell has a value denoting the weight between two nodes: A_{ij}=weight of the connection from i to j. The modularity for weighted networks can be calculated as (Clauset, Newman, & Moore, 2004):

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

m denotes the total number of links in the network: $\frac{1}{2}\sum_{ij} A_{ij}$. k_i is the degree of a vertex i in a weighted network: $\sum_j A_{ij}$. δ function is 1 if $u = v$ and 0 otherwise. The above formula is the fraction of within-community edges $\frac{1}{2m}\sum_{ij} A_{ij}\delta(c_i, c_j)$ minus the expected value of the same degrees of vertices randomly connected between the vertices: $k_i k_j/2m$, similar to unweighted networks.

The VOSviewer clustering technique was developed by Waltman, Eck, and Noyons (2010). It is a variant of Clauset, Newman, and Moore's (2004) community detection on weighted networks. The advantage of their method is that it unifies mapping and clustering approaches by solving the issue of minimizing:

$$V(x_1, \ldots, x_n) = \sum_{i<j} s_{ij} d_{ij}^2 - \sum_{i<j} d_{ij}$$

where $S_{ij} = \frac{2mA_{ij}}{k_i k_j}$ and d_{ij} has two options. For mapping, d_{ij} denotes the distance between nodes i and j in a p-dimensional map: $d_{ij} = \|x_i - x_j\| = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{ik})^2}$; for clustering, $d_{ij} = \begin{cases} 0 & \text{if } x_i = x_j \\ 1/\gamma & \text{if } x_i \neq x_j \end{cases}$ where γ is called the resolution parameter.

Waltman et al. (2010) have shown that minimizing V is equivalent to maximizing

$$\hat{V}(x_1, \ldots, x_n) = \frac{1}{2m}\sum_{i<j} \delta(x_i, x_j) w_{ij}\left(A_{ij} - \gamma\frac{k_i k_j}{2m}\right)$$

where w_{ij}=2m/k_ik_j. By examining *V hat* and *Q* for a weighted network, it can be found that *Q* is a special case when the resolution parameter γ and the weights w_{ij} are set equal to 1.

K-means and VOSviewer are applied to PPAM and PPWM. The clustering results for the two clustering methods are first compared, and the clustering results for the two networks are then compared and discussed.

Results

K-means for PPAM and PPWM

K-means is first applied to PPAM and PPWM. Choosing an appropriate k value is a delicate task for k-means algorithm. Cost function is used to calculate the sum of distances for k=5 to k=30 (see Figure 2).

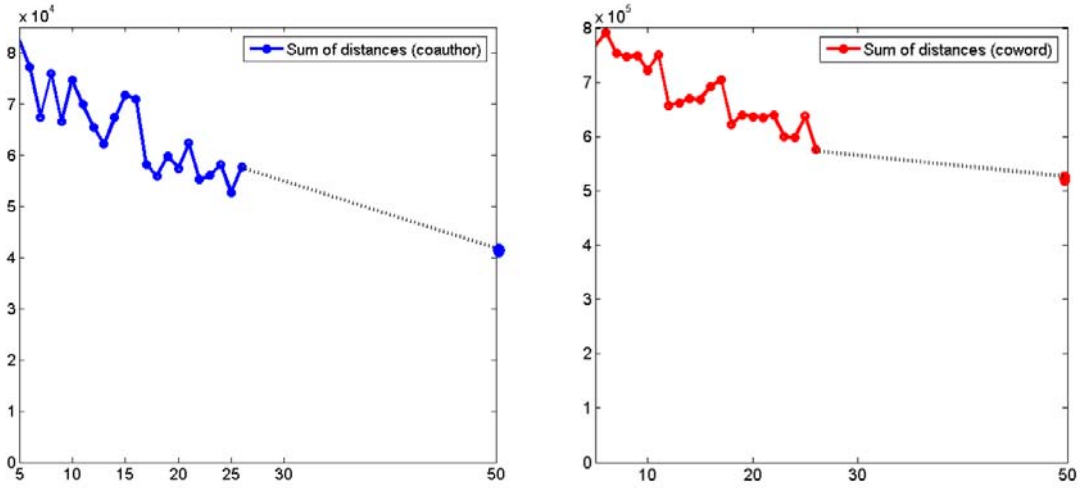

Figure 2. Sum of distance for k-means

As can be seen in Figure 2, the sum of distances are declining in a single direction, suggesting that choosing a relatively larger k would yield better clustering results. For comparison purposes, k was set to 50 for k-means because the VOSviewer clustering technique had identified approximately 50 clusters for PPAM and PPWM. The cluster sizes by k-means are displayed in Figure 3.

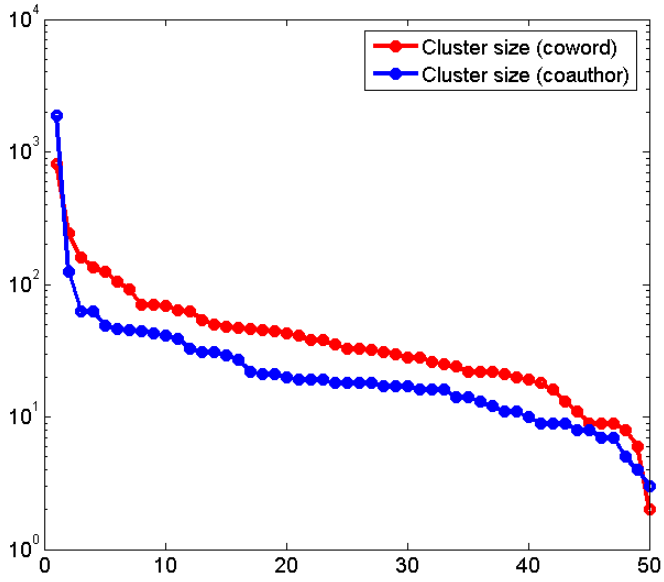

Figure 3. Cluster size (k-means)

The largest k-means cluster in PPAM contained more than 1800 authors and incorporated 60% of all nodes in PPAM; this result may not be a good partition as it failed to detect sub-groups from the largest cluster. Cluster sizes for PPWM were better allocated, with six clusters containing more than 100 nodes. The different clustering results for PPAM and PPWM may be attributed to the link densities of the two matrices: cell values in PPWM were generally larger than cell values in PPAM since two papers are more likely to share title words than authors (sum of total cell values in PPAM=76,317; sum of total cell values in PPWM=1,183,039). Therefore, it can be concluded that k-means works well for dense networks but is less effective with sparse networks.

VOSviewer for PPAM and PPWM

VOSviewer clustering technique was applied to both PPAM and PPWM. For the PPAM matrix, 54 clusters were identified; and 66 clusters were identified for PPWM matrix. The results can be found at: cluster_result.xlsx[4]. Cluster sizes range from 164 to 4 papers for PPAM and from 460 to 1 paper for PPWM (see Figure 4 for the distributions of cluster sizes). This indicates that the VOSviewer clustering technique was able to successfully detect smaller groups through limited links.

[4] http://info.slis.indiana.edu/~eyan/papers/cluster/cluster_results.xlsx

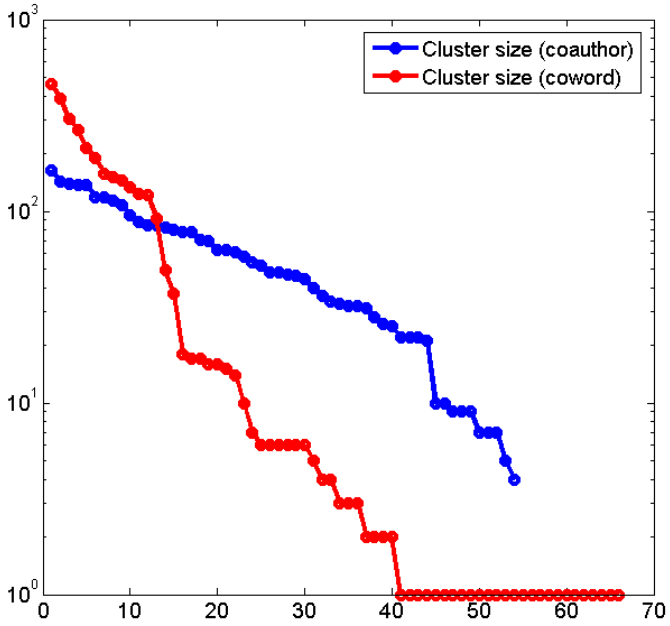

Figure 4. Cluster size (VOSviewer)

Dunbar (1998) predicted that roughly 150 members is the upper limit of a well-functioning human community. Several other studies have also found that smaller communities are desirable. For example, Allen (2004) found that on-line communities usually have 60 members and that a community will break into several smaller new communities if there are more than 80 members. Leskovec et al. (2008) found that communities greater than 100 nodes will gradually blend into the core of the network and thus become less community-like "with a roughly inverse relationship between community size and optimal community quality" (p. 1). Therefore, the clustering results from VOSviewer are preferred as most communities have approximately 100 nodes. The third research question is thus addressed.

PPAM: K-means vs. VOSviewer

Network visualizations are powered by VOSviewer. VOSviewer clustering technique was first applied to PPAM (see Figure 5). Using the same layout, k-means was then applied to PPAM (see Figure 6).

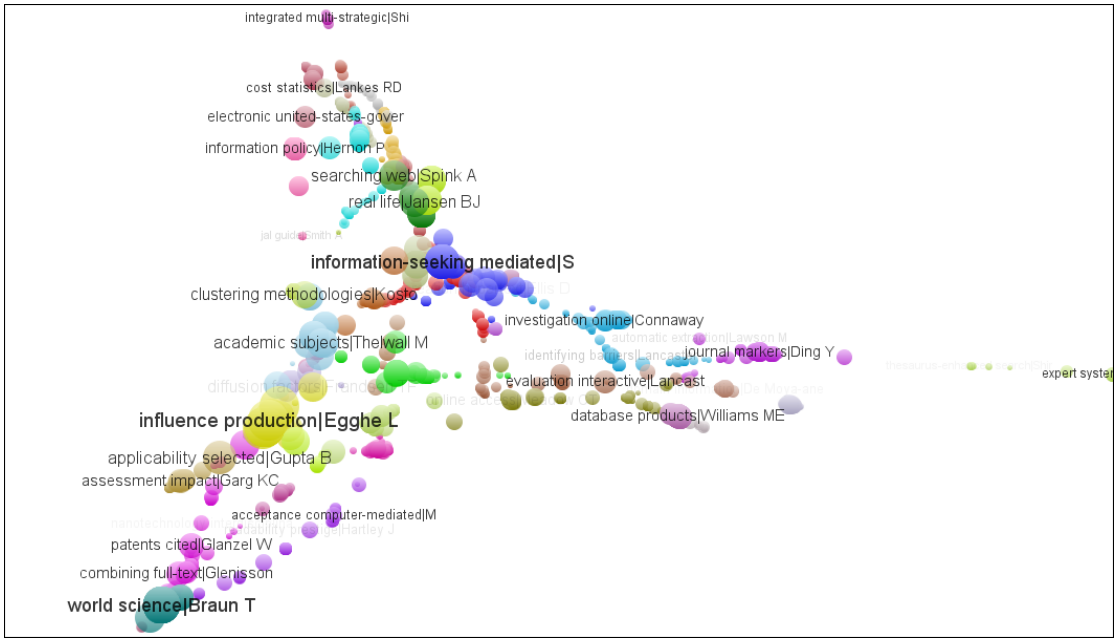

Figure 5. VOSviewer clustering visualization (PPAM)

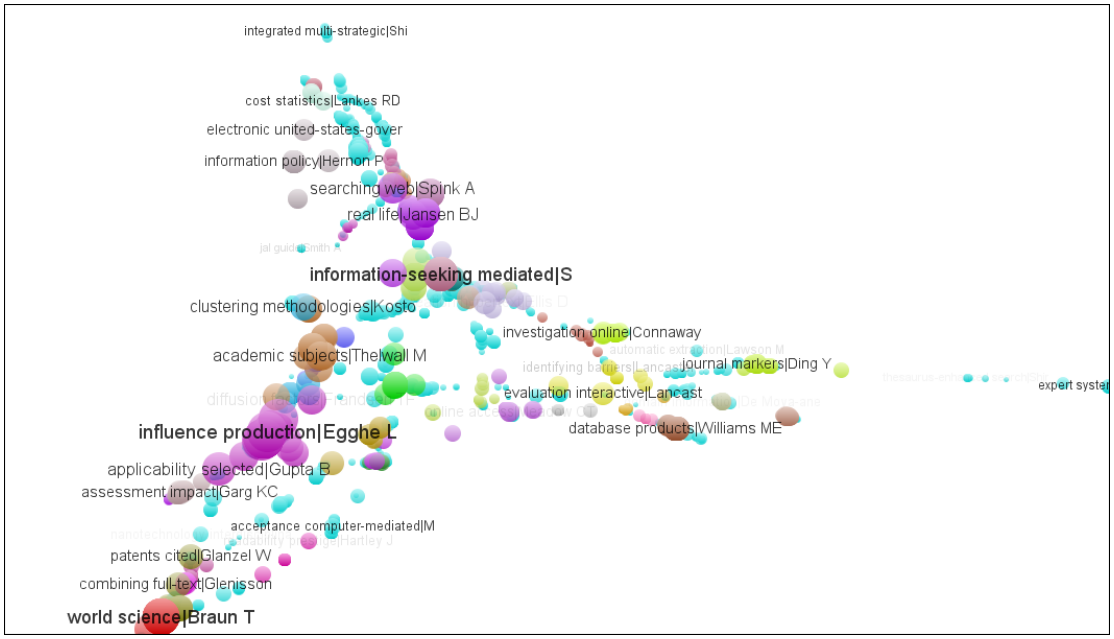

Figure 6. K-means clustering visualization (PPAM)

Comparing the two network visualizations In Figures 5 and 6, it can be seen that nodes in the largest cluster based on k-means (i.e., the bright blue bubbles) are scattered throughout the whole graph, suggesting that there are inconsistencies in the results for k-means and VOSviewer precisely because k-means did not divide the largest cluster into smaller clusters whose nodes were closer to each other.

PPWM: K-means vs. VOSviewer

VOSviewer clustering technique was then applied to PPWM (see Figure 7). Using the same layout, k-means was also applied to PPWM (see Figure 8).

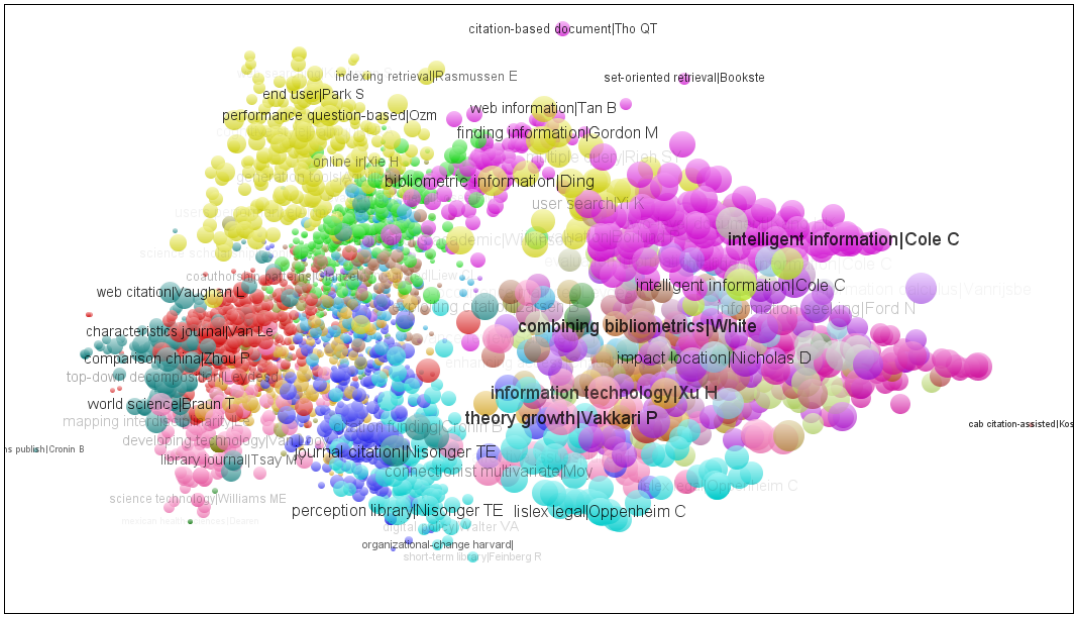

Figure 7. VOSviewer clustering visualization (PPWM)

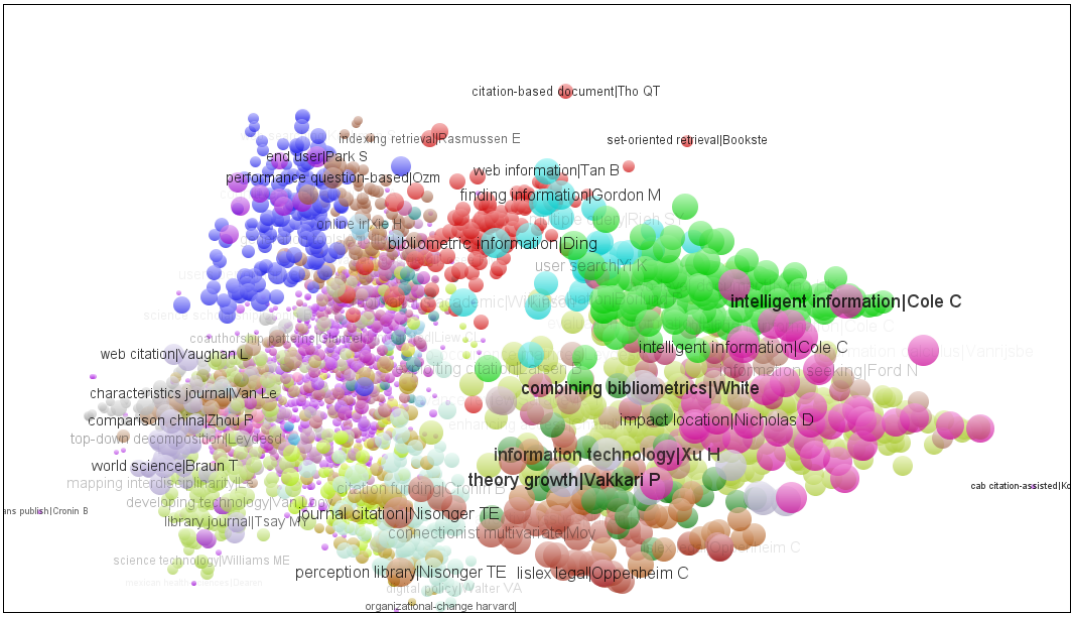

Figure 8. K-means clustering visualization (PPWM)

Comparing these two visualizations, it can be seen that k-means and VOSviewer clustering yield similar results in that nodes belonging to the same cluster are well collocated, indicating that k-means and VOSviewer are likely to yield similar results for dense networks.

Comparing shared author and shared word relations

Considering the ineffective clustering results on k-means, only clustering results on VOSviewer are used to compare shared author and shared word relations. As the cluster size distribution in Figure 4 indicates, cluster size follows a power law distribution. Therefore, small clusters were not chosen for comparing PPAM and PPWM as they would be likely to distort the matching results. For VOSviewer clustering, the top 31 clusters for PPAM and the top 12 clusters for PPWM were selected. Each of the selections covers more than 85% of the 3,053 publications. The 31 PPAM clusters were then matched with the 12 PPWM clusters to

form a 12*31 matrix where the cell value denotes the number of overlapping publications in the two partitions (i.e., on shared author relations and shared word relations).

For a more informative presentation of matching, a heat map was used. In the map depicted in Figure 9, cell values were normalized by dividing the product of row sum and column sum ($n_{ij}=c_{ij}/\sum_i c_{ij}*\sum_j c_{ij}$). Each row represents a particular research topic and the brightness of a cell indicates the dominance of a research community working on that topic; each column represents a separate research community, and the brightness of a cell represents how specialized a research topic is within that community.

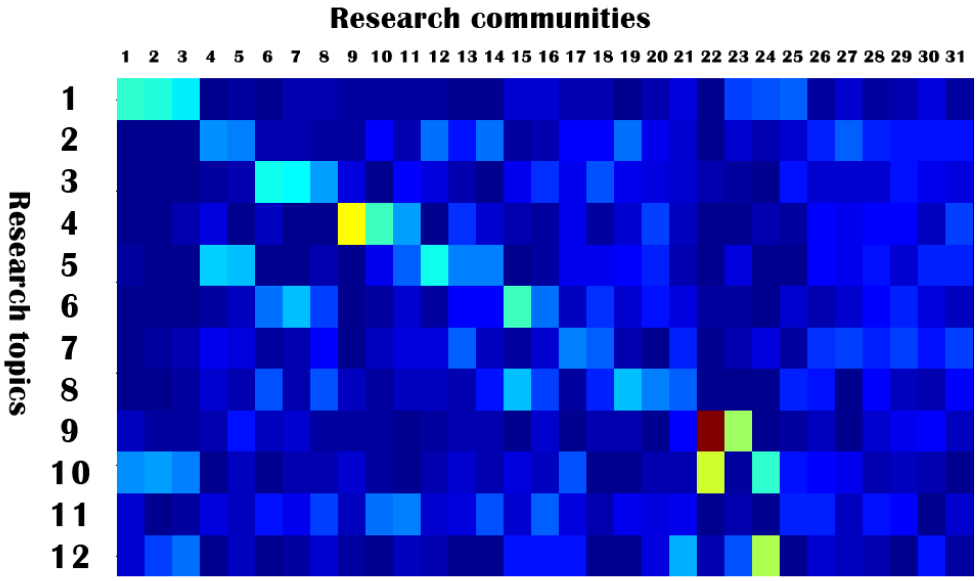

Figure 9. VOSviewer clustering result match

Figure 9 indicates that there are a few publication clusters (i.e., the cells depicted in brighter colors) that are highly specialized, with their authors writing on similar research topics and these research topics are worked on exclusively by these authors. Other publication clusters are less focused, with a given research topic studied by several different research communities and a single research community working on several research topics.

Analysis of publication titles and journals identified topics for 10 of the 12 PPWM clusters (see Table 2). These PPWM clusters can be further categorized into three broad topical areas: bibliometrics, information retrieval, and library science related topics. Two clusters (i.e., clusters 7 and 11) do not have a distinguishable topic, and this is evident in Figure 9, as well, where there is no brightly colored cell for either of these two rows. We find many titles of the articles in the two clusters have short titles and moreover common words in LIS were found (e.g. information, library, study, resource, analysis, etc.), resulted in the lack of topicality. These publications were clustered together since they are strongly connected internally (as they only share a few common words with each other) but loosely connected externally (as they do not share definite words with other publications). A majority of these publications would be "merged" and "absorbed" by other clusters if we tune down the resolution parameter γ (see Method section).

Table 2. Topics of clusters on PPWM

PPWM Clusters ID	*Cluster topics*
1	bibliometrics - scientific evaluation; bibliometrics - indicators

2	databases; information retrieval - classification; information retrieval - feedback; information retrieval - queries
3	libraries - digital libraries; libraries - reference; libraries - library evaluation
4	web - information seeking; web - webometrics
5	information retrieval - indexing
6	libraries - library collections; libraries - reference
7	-
8	scholarly communication
9	bibliometric - bibliometric laws; information retrieval - queries
10	information retrieval - queries
11	-
12	language processing

As can be seen in Table 2 and Figure 9, a few of the PPAM publication clusters have more focused concentrations, such as cluster 4 (web - webometrics) and 12 (language processing). These research topics are studied by authors who frequently co-occur in publications on these topics but less frequently in publications on other topics. Therefore, it can be noted that topic-wise, a few topics can be derived distinctively from research communities, thus answering the first research question.

There are also indications that a research topic may be studied by several research communities because of physical distance. For example, for the topic bibliometrics-scientific evaluation is studied by two separate research communities: one research community is comprised primarily of scholars from North America while the other is made up of scholars from Europe, the majority of whom are from the Netherlands and Hungary. Therefore, the formation of communities is affected by research topics; however, other factors, such as physical distance, also drive community structures. The second research question is thus addressed.

Compared with previous studies on research specialities and topics in LIS, the current study yields consistent results in that bibliometrics, information retrieval, and library science are the most visible research specialties in LIS. In addition, we also find that web related topics, such as webometrics and language processing (especially those targeting blogospheres), receive a growing attention in LIS.

Data used in co-citation analysis are derived from cited references: two cited authors will be connected by a tie if they are cited by a paper. Data used in paper-to-paper network analysis are directed harvested from citing articles: two papers will be connected by a tie if they share authors or title words. Although the two types of analyses use different data formats, the underlying assumption for both analyses is rooted in the idea of co-occurrence: the more elements two items share, the more similar the two items are. In addition, we argue that the paper-to-paper network may be a more accurate tool in mapping research specialties as clustering methods are directly applied to publications instead of authors, and thus there is no need to check each author's research expertise in order to determine the author's research specialties.

Conclusion

In this study, two layers of enriched information were constructed for communities: a paper-to-paper network based on shared author relations and a paper-to-paper network based on shared word relations.

K-means and VOSviewer, a modularity-based clustering technique, were used to identify publication clusters in the two networks. VOSviewer effectively partitioned PPAM, which had a limited number of links; in contrast, k-means was not able to detect sub-groups in the

largest PPAM cluster, indicating that k-means is not as effective when applied to sparse networks.

By matching the clustering results for PPAM and PPWM, it was found that many of the communities had a relatively distinct research specialty: authors who shared similar research interests tended to work together and, as a result, they published articles on similar research topics such that topics could be derived from analysis of the research community itself. Furthermore, a few research topics (e.g., webometrics and language processing) were uniquely associated with a specific research community while other research topics (e.g., information retrieval-queries and information retrieval-indexing) were studied by several different research communities. This demonstrates that the composition of a research community is frequently driven, at least in part, by the topic of study, even though other factors, such as physical distance, may also play a role in community development.

The results of this study indicate that future research on this topic would benefit from adding dynamics to research communities in order to determine how topics interact with communities and how communities may co-evolve with the topics they research.

Acknowledgement

The authors would like to thank Ludo Waltman, Nees van Eck, and Ed Noyons for introducing VOSviewer and their comments to the idea of this paper.

References

Allen, C. (2004). Life with alacrity: The Dunbar number as a limit to group sizes. Retrieved April 19, 2010 from http://www.lifewithalacrity.com/2004/03/the_dunbar_numb.html

Åström, F. (2010). The visibility of information science and library science research in bibliometric mapping of the LIS field. *The Library Quarterly, 80*(2), 143-159.

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

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large network. *Physical Review E, 70*, 066111.

Ding, Y. (2011 submitted). Community detection: Topological vs. topical. Manuscript submitted to *Journal of Informetrics*.

Donetti, L., & Munoz, M. A. (2004). Detecting network communities: a new systematic and efficient algorithm. Journal of Statistical Mechanics: Theory and Experiment, 2004, P10012.

Dunbar, R. (1998). *Grooming, gossip, and the evolution of language*. Boston, MA: Harvard University Press.

Farkas, I., Ábel, D., Palla, G., & Vicsek, T. (2007). Weighted network modules. *New Journal of Physics, 9*(6), 180-209.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports, 486*(3-5), 75-174.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America, 99*(12), 7821–7826.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 50-57, Aug 15-19, 1999, Berkeley, CA, USA.

Leskovec, J., Lang, K. J., Dasgupta, A., & Mahoney, M. W. (2008). Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. Retrieved March 21, 2010 from http://arxiv.org/abs/0810.1355

Li, D., He, B., Ding, Y., Tang, J., Sugimoto, C., Qin, Z., Yan, E., & Li, J. (2010). Community-based topic modeling for social tagging. *The 19th ACM International Conference on Information and Knowledge Management (CIKM2010)*, pp: 1565-1568, Oct 26-30, 2010, Toronto, Canada.

Morris, S. A., Van der Veer Martens, B. (2008). Modeling and mapping of research specialties. *Annual Review of Information Science and Technology, 42*(1), 213-295.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences, 103*(23), 8577-8582.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences, 98* (2), 404-409.

Newman, M., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*, 026113.

Nisonger, T.E. & Davis, C. H. The perception of library and information science journals by LIS education deans and ARL library directors: A replication of the Kohl-Davis study. *College & Research Libraries, 66*, 341-77.

Persson, O. The Intellectual base and research fronts of JASIS 1986-1990. *Journal of the American Society for Information Science, 45*(1), 31-38.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America, 101*(9), 2658-2663.

Richardson, T., Mucha, P. J., & Porter, M. A. (2009). Spectral tripartitioning of networks. *Physical Review E, 80*, 036111.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 306-315, New York: ACM Press.

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*, pp.990-998.

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

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

White, H. D., & Mccain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science, 49*(4), 327-355.

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