Scientometrics (2014) 101:309–335
DOI 10.1007/s11192-014-1417-1

Investigating the integrated landscape of the intellectual topology of bioinformatics

Meen Chul Kim · Yoo Kyung Jeong · Min Song

Received: 20 August 2013 / Published online: 6 September 2014

Abstract We aim at identifying (1) whether and how various data sources influence mapping an intellectual structure of the field of bioinformatics, and (2) the landscape of bioinformatics by integrating those sources. To this end, we conduct a comprehensive bibliometric analysis by harvesting bibliographic information from DBLP, PubMed Central, and Web of Science. We then measure and compare topological characteristics of networks generated using these sources. The results show a dichotomous pattern dominated by PubMed Central and WoS. In addition, a few influential scientists in the field of bioinformatics receive very high citations from their colleagues, which is a driving force to bloom the field. These few scientists are connected to a much larger research community. Most of the researchers are intellectually linked within a few steps, in spite of the domain's interdisciplinary characteristics. Particularly, influential authors consist of a small world. We also identify that there is not a coherent body of discipline in bioinformatics since the field is still under development. Finally, the journals and conferences indexed by each source cover different research topics, and PubMed Central is more inclusive than DBLP as an indexing database.

Keywords Bibliometrics · Big data · Bioinformatics · Complete topology · Co-authorship network · Co-citation network

Introduction

Bibliometrics is a research method which uses quantitative and statistical analyses to describe patterns of publication within given field or body of literature. Researchers use

M. C. Kim
College of Computing and Informatics, Drexel University, Philadelphia, PA, USA
e-mail: meenchul.kim@drexel.edu

Y. K. Jeong · M. Song (✉)
Department of Library and Information Science, Yonsei University, Seoul, Republic of Korea
e-mail: min.song@yonsei.ac.kr

Y. K. Jeong
e-mail: yk.jeong@yonsei.ac.kr

bibliometrics to determine the influence of a certain scientist, or to describe the relationship between two or more researchers or works. In this respect, bibliometrics shares similar scientific objectives with social network analysis (SNA) which focuses on discovering implicit structures in the social environment in which certain relational and interactive units exist (Wasserman and Katherine 1994). In SNA, the units of analysis are various actors (users of social web) and topics (e.g. opinions, diseases) whereas in bibliometrics the units are authors and bibliographic entities such as papers and journals.

Both bibliometrics and SNA have attracted interest from various scientific communities over the last decades. Bibliometric analyses reveal interesting features of academic and social communities; they can be used to uncover cohesive collaboration among researchers and invisible communities, and to represent the intellectual structure of a knowledge domain (Newman 2001). In addition, these approaches often reflect the movement of an author's research domain (Huang and Huang 2006). SNA views social relationships in terms of network theory, consisting of nodes representing individual actors within the network and ties which represent relationships between the individuals, such as friendship, kinship, organizations, sexual relationships, etc. (Abraham et al. 2010). It employs graph theory which focuses on identifying mathematical structures used to model pairwise relations between objects.

As mentioned earlier, both bibliometric and scientific networks have been well-studied (e.g. Newman 2001, 2004; Velden et al. 2009). Some of these previous studies have aimed at visualizing datasets with the goal of highlighting the most influential researchers in academic domains and their relationships among them. Others have tried to uncover some statistical features by comparing citation databases (Kulkarni et al. 2009). Specifically in the field of bioinformatics, several researchers have applied bibliometric analyses to understand the development of the field (Glänzel et al. 2009; Huang et al. 2012; Janssens et al. 2007; Manoharan et al. 2011; Song et al. 2013a, b). A number of researchers have concentrated on not only content and author network analyses but biomedical entities as a unit of analysis for automatic discovery, investigation of protein–protein or drug–drug interactions, and so on (He et al. 2011; Kolchinsky et al. 2010; Marques-Pita and Rocha 2013).

Previous approaches, however, have a few limitations. First, they have relied on bibliographic information provided by an individual database (Huang et al. 2012; Manoharan et al. 2011; Patra and Mishra 2006; Song et al. 2013a) or they have selectively chosen core literatures (Glänzel et al. 2009; Janssens et al. 2007). Such silos of data sources prevent them from adequately exploring how a variety of characteristics derived from their own scientific objectives affect the formation of the intellectual structure. Thus, they have not considered a complete picture of bioinformatics. To our best knowledge, the present study is the first to integrate diverse data sources for bibliometric analysis in bioinformatics. Although some of the previous studies applied bibliometric analyses to bibliographic data of bioinformatics, their primary focuses were not on investigating the overall intellectual structure in the field of bioinformatics.

Our research objectives are three-fold. The first is to investigate whether and how various data sources (DBLP, PubMed Central, and Web of Science), which have different scientific scope in terms of indexed journals and conference papers, influence mapping the intellectual structure of bioinformatics. The second is to identify a composite landscape of bioinformatics by integrating the bibliometric networks derived from those sources if they demonstrate only a partial picture. The third is to apply SNA techniques to analyze the result produced by bibliometric methods to examine if there are distinctive characteristics of bioinformatics communities. We employ two different approaches of bibliometric analysis: (1) co-authorship analysis, and (2) co-citation analysis.

The next section discusses related work. We introduce the procedure for co-authorship and co-citation network analyses in the "Methodology" section. We compare and analyze the experiment in the "Results" section.

Related works

There have been rigorous attempts to discover intellectual structures that are representative of selected knowledge domains. They empirically have the existence of scientific networks. A bibliometric network is generally defined as a set of vertices (relationships) between nodes (authors), with some additional information on the vertices and/or the nodes in the graph (Barnett 2011). That is, the graph's nodes represent academic researchers and the graph's edges represent scientific interactions (Nooy et al. 2005). In addition, they offer techniques and concepts to describe structural properties of networks such as cohesive subgroups, brokerage roles of the members of networks and detecting potential hierarchies in networks (Nooy et al. 2005; Wasserman and Katherine 1994).

Numerous researchers have applied bibliometric analyses to reveal scientific interactions among researchers in research domains including information science and bioinformatics. In the following section, we discuss previous studies.

We focus on two major approaches to bibliometric analysis, co-authorship and co-citation analyses. Co-authorship analysis focuses on an author's collaborators as a unit of analysis, while co-citation analysis is based on a pair of authors cited by a third party. We adopt both perspectives for scientific network analysis in this paper.

Bibliometric approaches in the field of bioinformatics

Bibliometric analyses have been used to identify scientific interactions among bioinformatics researchers, and to investigate the formation and evolution of the field itself. Some papers have explored the rapid and diversified growth of the scientific literatures in bioinformatics (Patra and Mishra 2006; Perez-Iratxeta et al. 2007; Song et al. 2013b). Those studies have identified key features of the field such as core primary journals and productivity patterns of authors and their institutions by analyzing literatures indexed in Web of Science or PubMed Central. In addition, sub-areas of rapid growth were observed (Perez-Iratxeta et al. 2007). In their analysis of a few key journals indexed by DBLP, Song et al. (2013a) found an increasing overlap among bioinformatics journals in terms of topics based on their measurement of content and network similarities. Huang et al. (2012) and Manoharan et al. (2011) conducted bibliometric experiments covering manuscripts indexed by Web of Science. Through these studies, they identified that just a few countries tend to produce the majority of the publications in the domain (Manoharan et al. 2011), and that citations of the bioinformatics journals were field-dependent with scattered patterns in article life span and citing propensity (Huang et al. 2012). Based on 3,910 publications by Brazilian authors on seven diseases retrieved from Web of Science, Morel et al. (2009) identified co-authorship network maps with selected biomedical terms (disease). Finally, in addition to citation-based bibliometric techniques, some studies have instead aimed to enhance the performance of text mining techniques for network analysis (Glänzel et al. 2009; Janssens et al. 2007).

However, these approaches are based on bibliographic information provided by individual data sources or they selectively chose core literatures. None of these considered how various data sources which have different scientific scope in terms of indexed journals and conference papers may have influenced the outcome of mapping the intellectual structure of a certain domain. Therefore, this study focuses on identifying the characteristics of various data sources, each with their own academic scopes, and how they affect the depiction of the intellectual structure of a certain domain when analyzed using bibliometric techniques. Ultimately, we show a comprehensive and robust network for the field of bioinformatics.

Co-authorship analysis

Scientific collaboration is a complex social phenomenon that has been systematically studied (Glänzel and Schubert 2004). In early studies, many researchers tried to identify why and how collaboration occurs. Several studies reported increased collaboration among researchers in the context of growing funding, that is, collaboration is affected by economic factors (Clarke 1964, 1967; Heffner 1981; Price and Beaver 1966; Smith 1958). It was later found that these factors motivate co-operation in 'less expensive' fields such as pure mathematics and theoretical research in social sciences (Glänzel and Schubert 2004). Besides the economic factors, intra-scientific factors, especially changing communication patterns and increasing mobility of scientists, have also been shown to stimulate collaboration (Beaver and Rosen 1978, 1979; Luukkonen et al. 1992, 1993). Subsequent research focused on discovering how co-authorship relates to scientific communities in a specific research domain. In this context, some research has performed small-scale statistical analyses of such as frequency of co-authored articles by particular authors or authors at particular institutions (Newman 2001).

With the advent of comprehensive online bibliographies, the construction of complete or near-complete co-authorship networks for entire fields became possible. Drawing in these new capabilities, researchers established large-scale networks representing research in nearly every domain including biology, computer science, and physics. Their findings confirmed the existence of the invisible community in analyzing topologies, showing the distinct differences between author clusters. These works also showed (1) intellectual maps of specific scientific areas (Day et al. 2010; Erma and Todorovski 2010; Glänzel and Schubert 2004; Ioannidis 2008; Morel et al. 2009), (2) sub-clusters representing further detail of the collaborative structure of particular areas (Hou et al. 2006), and (3) academic and communicative practices of certain domains (Catala-Lopez et al. 2012; Day et al. 2010; Glänzel and Schubert 2004; Newman 2001; Velden et al. 2009). Other studies have compared co-authorship networks among different scientific fields, and analyzed collaborative practices in domain-dependent communities (Newman 2001, 2004). These results reveal that each scientific community seems to constitute its own 'small world', meaning the networks are highly clustered. Other works weigh an individual's contribution in a co-authored work (Hany et al. 2009; Huang and Huang 2006), and applied visualization techniques (Catala-Lopez et al. 2012; Huang and Huang 2006). A co-authorship network analysis in bioinformatics was reported by Song et al. (2013a). They used selected journals and conferences from DBLP for the co-authorship analysis and found that bioinformatics is fast-growing, dynamic and field dependent. By analyzing and comparing co-authorship networks from three different databases, we seek to examine how the characteristics of each data source affect the conceptualization of an invisible communities.

Co-citation analysis

Knowledge flow may be seen as a sociocultural phenomenon in which knowledge is transferred from one to another entity among people, friends, families, communities, or organizations. Such transfers have been investigated, based on links from cited papers to citing papers. A citation analysis which involves pairs of authors who are cited together is called an "author co-citation" analysis (Kim and Barnett 2008). Traditionally, the goal of co-citation analysis is to identify whether cited papers are conceptually connected through a citing paper and how many co-citations a paper has accumulated. White and Griffith (1981) introduced the concept of co-citation analysis as a literature measure for observation of intellectual structure. This foundational study visualized the intellectual structure of the field of information science, and the resulting map showed identifiable author groups, proximities of authors within group and across group boundaries, and positions of authors using a Multi-Dimensional Scaling (MDS) space. Several subsequent studies have revealed (1) research trends in diverse knowledge domains, (2) formation, migration, and diffusion of information, and (3) characteristics of scholarly communication of particular scientific domains (Perry and Rice 1998; White 2003b). While many studies in this area have applied the general steps and techniques of classic author co-citation analysis to different research domains with minor or few modifications, a few have focused on the application of different statistical approaches to investigating various characteristics of co-citation networks (Ahlgren et al. 2003; Kulkarni et al. 2009; Leydesdorff 2005; Leydesdorff and Vaughan 2006; White 2003a). For instance, there is disagreement as to whether a correlation measure is appropriate for analyzing the co-occurring authors appearing on co-citation count matrices, Other studies have focused on demonstrating different approaches to counting frequency of co-citations such as first author co-citation and pure co-citation (Persson 2001; Rousseau and Zuccala 2004; Zhao 2006).

Co-citation analysis has been employed in conjunction with SNA techniques. Wasserman and Katherine (1994) argued that citation analysis might be a plausible approach of the social network studies that use bibliographic data on "who cites whom". Other researchers explored the relationship between the concepts from SNA and bibliometrics (Ding et al. 2009; White et al. 2004). The latter literature investigated whether citers and citees have interpersonal as well as intellectual ties. The result of the study identified co-citation as a powerful predictor of those relationships. Ding et al. (2009) employed the PageRank algorithm (Brin and Page 1998), to explore co-citation networks. They showed that citation rank is highly correlated with PageRank.

Methodology

Figure 1 shows our citation analysis procedure. Detailed descriptions are provided in the following sections.

We collected bibliographic information about bioinformatics from three different sources, Web of Science, DBLP, and PubMed Central. The datasets from DBLP and PubMed Central data sets are in XML format, and Web of Science provides its data as a CSV spreadsheet file. We developed XML and spreadsheet parsers to extract needed bibliographic elements such as author, title, and year. After that, we created co-author and co-citation pairs. In the last phase, we built co-authorship and co-citation networks for analysis.

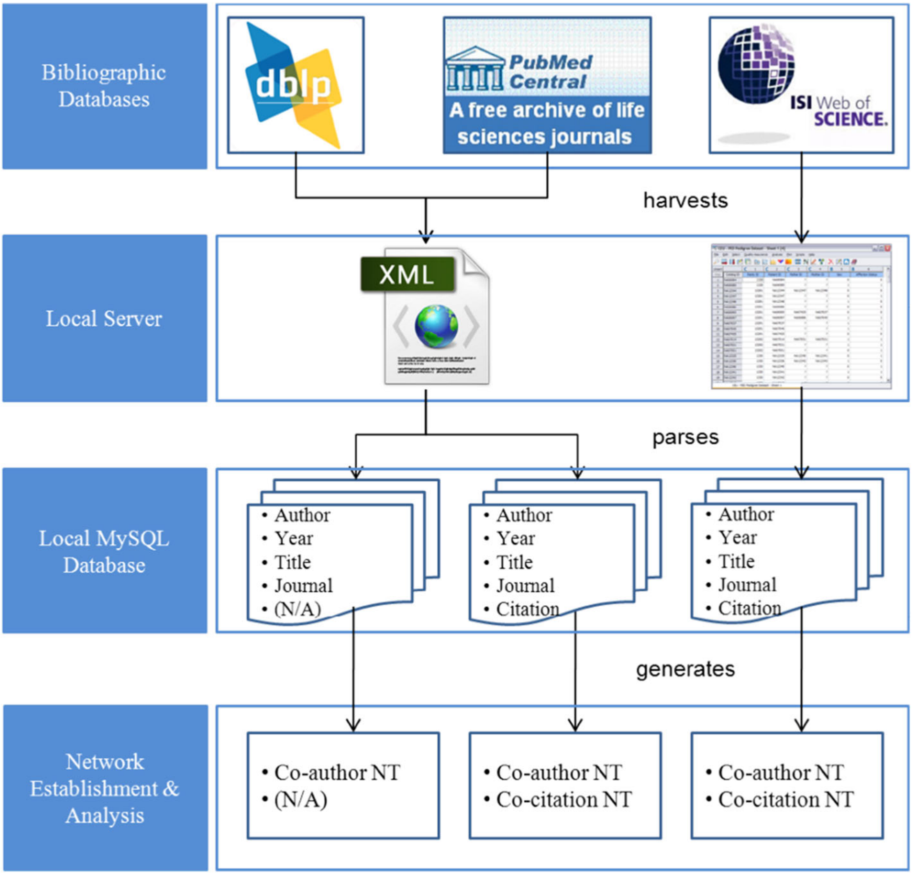

Fig. 1 Citation analysis procedure

Data collection

As argued earlier, we aim to identify more complete landscapes of bioinformatics by author collaboration and author co-citation analysis. Nonetheless, the distinction of the domain remains unclear because the field of bioinformatics is highly interdisciplinary and relatively new. We operationalize bioinformatics as follows: conceptualizing biology in terms of macromolecules (in the sense of physical-chemistry) and then applying "informatics" techniques (derived from disciplines such as applied maths, computer science, and statistics) to understand and organize the information associated with these molecules, on a large-scale (Luscombe et al. 2001). This definition rationalizes our criteria of collecting datasets.

To accomplish our research objectives with the above definition, we collected data from the three different databases which index scientific papers in the field of bioinformatics. Two important criteria for the database selection were: (1) whether the subject matter of the databases covered the field of bioinformatics, and (2) whether the databases had complete bibliographies to index papers. It was important to gather several data sets since large-scale data collection influences the construction of complete or near-complete co-authorship/co-citation networks. Given these criteria, we chose three databases—DBLP, PubMed Central, and ISI Web of Science.

DBLP is one of the largest freely available bibliographic data sources on Information Science. It provides information on major computing journals and conference proceedings between the years 1936 and 2012. It also tracks proceedings papers of many conferences. From DBLP, we collected bibliographic information of 14,604 papers from 52 indexed conference proceedings and journals in the field of bioinformatics, from 1985 to 2012 inclusively. The data included (1) 15 conferences classified as bioinformatics conferences by DBLP (http://www.informatik.uni-trier.de/~ley/db/bio.html) and (2) 97 conferences and journals (http://www.informatik.uni-trier.de/~ley/db/conf/indexa.html) that were selected if the term "bio" were part of the conference name in the DBLP category. After removing duplicates from these two sources and irrelevant conferences to bioinformatics such as Bio-Inspired Design of Networks (BIOWIRE), International Conference on Biology, Informatics, and Mathematics (JOBIM), Languages in Biology and Medicine (LBM), Knowledge Discovery and Emergent Complexity in Bioinformatics (KDECB), and Genetic and Evolutionary Computing (WGEC), we ended up including 52 conferences and journals for analysis. The bibliographic information from DBLP, however, does not include citations. Therefore, we only include this dataset in co-authorship analysis but exclude it from co-citation analysis.

PubMed Central is a database of full-text scientific literature in biomedical and life sciences developed by the US National Library of Medicine (NLM) as an online archive of biomedical journal articles. As of January 2013, the archive contains approximately 2.6 million full-text items, including articles, editorials and letters. For PubMed Central, we crawled bibliographic data of 20,869 articles from 49 indexed journals in the same field, from 2001 to 2012 inclusively. We used the same selection criteria as Song et al. (2013b).

Web of Science (WoS) is an online subscription-based scientific citation indexing service maintained by Thomson Reuters that provides a comprehensive citation search. It gives access to multiple databases that reference cross-disciplinary research, which allows for in-depth exploration of specialized sub-fields within an academic or scientific discipline. We downloaded 62,523 records indexed in 47 journals in the field of bionformatics, from 1990 to 2012 inclusively. We selected the journals from the list of the Mathematical and Computational Biology section in Web of Science's Journal Citation Reports (JCR). Huang et al. (2012) and Song et al. (2013b) used the data from Web of Science's JCR section related in bioinformatics, and we also adopted this data selection criterion.

By basing our selection criteria justified by previous studies, we aim to unbiasedly collect journals and conference papers in the field of bioinformatics. We assume that collecting data from different sources would be beneficial to captivating a more complete topology in sciences. Therefore, our data selection criteria support our overarching research goal of identifying the complete landscape of bioinformatics.

We compiled our data set with author, year of publication, article title, journal title, and list of cited authors (except for DBLP as explained above). On average, there were 3.61 (DBLP), 5.67 (PubMed), and 3.61 (WoS) collaborative authors participated in a paper. The difference across databases might be originated from several reasons such as each database's indexing policy or scientific scope in collecting manuscripts. We aim to address this issue in the future study. In addition, PubMed Central has 32.08, and WoS has 30.30 cited references on average. Table 1 briefly summarizes our data collection.

Figure 2 shows the number of overlapping journals among databases. Only five journals are included in all three databases: Algorithms for Molecular Biology, Bioinformatics, BMC Bioinformatics, Evolutionary Bioinformatics, and PLoS Computational Biology. In

Table 1 Descriptive data statistics

	DBLP	PubMed Central	Web of Science
# of indexed venues	52	49	47
# of articles	14,604	20,869	62,523
# of authors	48,924	71,294	87,138
Average # of authors	3.35	3.42	1.39
Average # of co-authors	3.61	5.67	3.61
Average # of citations	N/A	32.08	30.30
Time span	1985–2012	2001–2012	1990–2012

addition, the overlap between pairs of databases also is low (DBLP and PubMed Central: 9; DBLP and Web of Science: 7; PubMed Central and Web of Science: 5). This supports the argument of our study that suggests that the inclusion of multiple databases will aid in establishing a complete map of bioinformatics. Through this triangulation of the dataset, we improve the data integrity. In addition, we assume that the few overlaps of journals and conferences across databases might be caused by multidisciplinarity and immaturity of the domain. Bioinformatics is relatively new, fast-growing, and co-evolving across the sub-communities (Song et al. 2013b). Therefore, it is likely that there is not a coherent field of bioinformatics.

Figure 3 illustrates that the entire number of articles per year indexed by two databases (Web of Science and PubMed Central) has increased exponentially since 2000s whereas DBLP shows a linear increase over full length of the period of analysis. This difference primarily comes from the fact that DBLP majorly provides bibliographic information on computer science journals and proceedings. Overall, these trends indicate the continual explosion of publications in this domain.

Establishing co-authorship and co-citation networks

Prior to extracting author-based networks, the problem of ambiguous author names needs to be addressed so that one node on each topology represents a different researcher. This is a challenging issue which cannot be addressed fully within the scope of the current paper; however, based on the characteristics of each dataset, we take different approaches to address this issue. First, PubMed Central offers completed author information including organizational affiliation. Therefore, we utilize both names and affiliations to identify unique authors. However, bibliographic descriptions from DBLP and Web of Science do not include an author's institutional affiliation. To disambiguate these data, we replace punctuation marks in the author field with white spaces, and make the author's name upper-case.

To construct co-author and co-citation pairs for author-based networks, we employed an intuitive approach. Given a paper p consisting of n number of collaborative authors (a_1, a_2,…, a_n) and cited articles (c_1, c_2,…, c_n), let each co-author a_n in a publication p be paired with other collaborators. Likewise, generation of co-citation pairs applies the same way to co-authors $c_n a_n$ in a cited article. For instance, assume that a certain article has four co-authors—a_1, a_2, a_3, and a_4. Then, six pairs can be generated as follows: (a_1;a_2), (a_1;a_3), (a_1;a_4), (a_2;a_3), (a_2;a_4), (a_3;a_4). By counting these cumulative author pairs, we measure the strength of the undirected vertices (ties between authors). For visualization of each

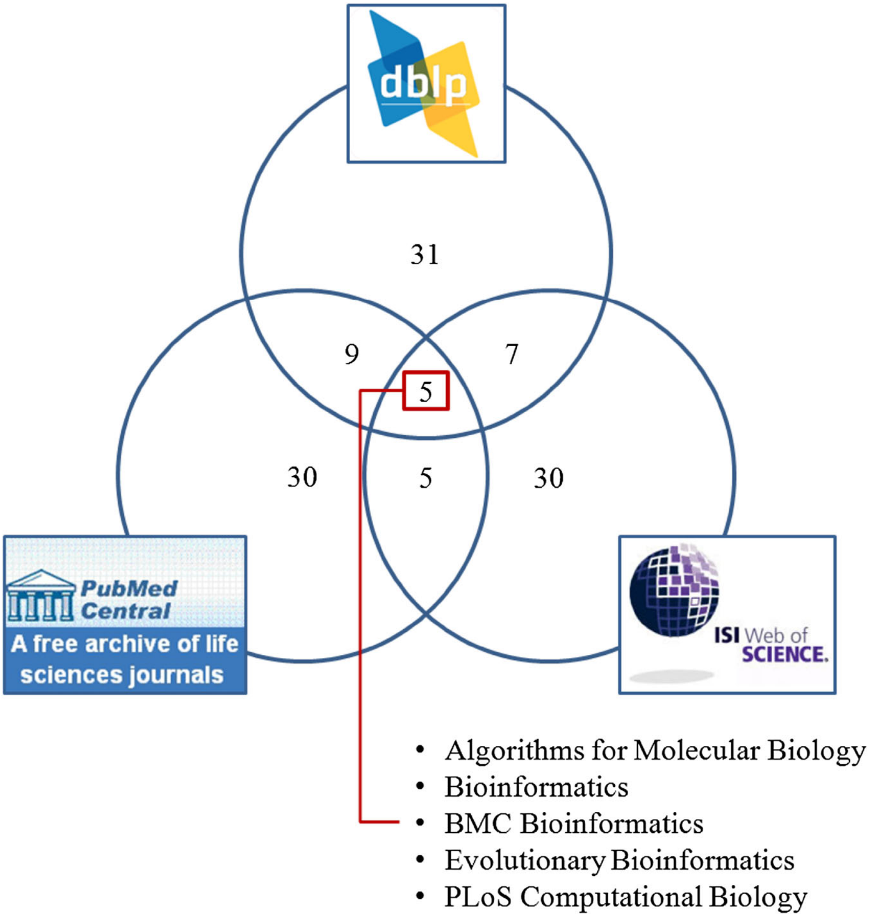
dblp
31
9
7
5
30
5
30
PubMed Central
A free archive of life sciences journals
ISI Web of SCIENCE.
• Algorithms for Molecular Biology
• Bioinformatics
• BMC Bioinformatics
• Evolutionary Bioinformatics
• PLoS Computational Biology

Fig. 2 Overlapping journals among three data sources

network, we employ Gephi, an open source visualization and manipulation tool (http://gephi.org/). Figure 4 describes the way we created these kinds of networks.

Topological measurement

In addition to establishing co-authorship and co-citation networks, we utilize their topology to compare methods of network formation using the following measures:

- *Degree centrality* is defined as the number of ties that a node has (Sade 1989). For a standardized score, each score needs to be divided by $n - 1$ where n refers to the number of nodes.
- *Clustering coefficient* is a measure of degree to which nodes in a graph tend to cluster together. The clustering coefficient for the whole network is the average of the local clustering coefficients of all the vertices n (Watts and Strogatz 1998).
- *Network diameter* is the number of shortest paths between any two nodes in the network. The diameter is representative of the linear size of a network (Chung 1984).

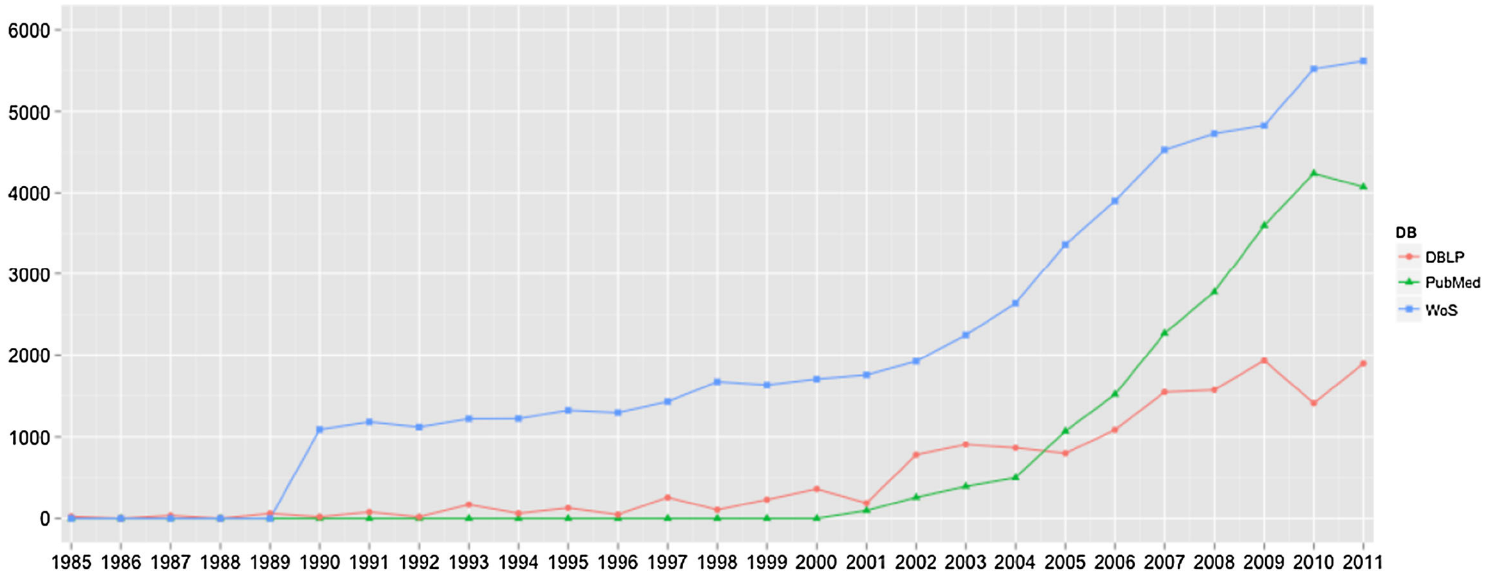

Fig. 3 Temporal comparison among the number of articles per year published by the journals and the conferences chosen

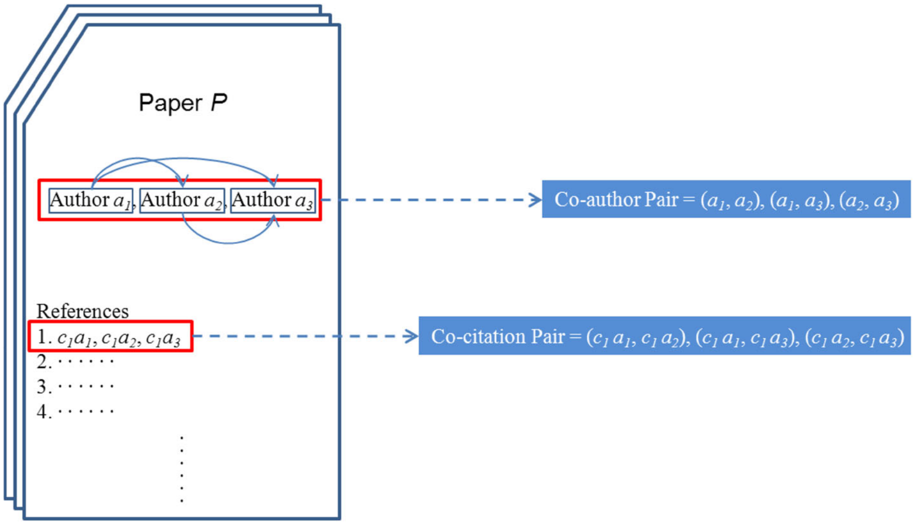
Paper P
Author a_1, Author a_2, Author a_3
Co-author Pair = $(a_1, a_2), (a_1, a_3), (a_2, a_3)$
References
1. c_1a_1, c_1a_2, c_1a_3
2. ······
3. ······
4. ······
Co-citation Pair = $(c_1 a_1, c_1 a_2), (c_1 a_1, c_1 a_3), (c_1 a_2, c_1 a_3)$

Fig. 4 Author network establishment

- *Graph density* is an indicator to measure how close the number of edges is to the maximal number of edges. If a certain graph has a few edges, it is a sparse graph (Coleman and Moré 1983).
- *Average path length* is defined as the average distance between all pair of its nodes (Chen et al. 2008). That is, it indicates the average number of steps along the shortest paths for all pairs of network nodes. It measures the efficiency of information or mass transport on a network.
- A community is a densely connected subset of nodes that is only sparsely linked to the remaining network. We apply *the modularity algorithm* (Blondel et al. 2008) to the entire network for discovering hidden behavioral or functional communities of the network. This technique also supports better exploration and browsing tools for very large collections.

Table 2 shows formula and value range of each measure criterion.

Results

Co-authorship analysis

First, we generated three co-authorship topologies from each dataset based on the aforementioned approach to network construction. Because of appropriateness of their sizes, we did not reduce the scale of each network. As a result, each network is built as follows: the DBLP network with 27,396 nodes and 68,865 edges, the PubMed Central network with 68,531 nodes and 466,614 edges, and the WoS network with 108,520 nodes and 322,140 edges. Table 3 specifies the statistical description of each co-authorship network.

Figure 5 shows the co-authorship networks. As suggested by comparisons with Table 3, these maps reflect the characteristics of each domain derived from the databases, and show distinct topological features per database. The researchers whose collaborative studies

Table 2 Measurements of network analysis

Criterion	Formula	Value range
Degree centrality	$C_D(G) = \frac{\sum_{i=1}^{\lvert V\rvert}[C_D(v*) - C_D(v_i)]}{H}$	N/A
Clustering coefficient	$C_n = \frac{2e_n}{[k_n(k_n - 1)]}$	Min. 0 Max. 1
Network diameter	$D = \frac{(n+1)}{3}$	Min. 1
Graph density	$D = \frac{2\lvert E\rvert}{\lvert V\rvert(\lvert V\rvert - 1)}$	Min. 0 Max. 1
Average path length	$L_G = \frac{1}{n(\mathrm{n}-1)}\sum_{i,j} d(v_i, v_j)$	N/A
Modularity	$\mathit{\Delta}\mathrm{Q} = \left[\frac{\sum_{\mathrm{in}} + 2k_{i,in}}{2m} - \left(\frac{\sum_{\mathrm{tot}} k_i}{2m}\right)^2\right] - \left[\frac{\sum_{\mathrm{in}}}{2m} - \left(\frac{\sum_{\mathrm{tot}}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$	Min. −0.5 Max. 1

Table 3 Co-authorship network statistics

	DBLP	PubMed Central	Web of Science
# of nodes	27,396	68,531	108,520
# of edges	68,865	466,614	322,140
# of components (sub-clusters)	2,893	3,565	9,269
Nodes in the giant component	14,041 (51.2 %)	49,601 (72.3 %)	69,374 (63.9 %)
Clustering coefficient	0.784	0.864	0.758
Network diameter	25	18	27
Graph density	0.000186	0.000198	0.000054
Average path length	9.561	6.253	9.365

The giant component represents the large group of people connected to one another through paths in the network (Kumar et al. 2010)

were published in the journals indexed by PubMed Central consist of the smallest world, based on the observation on the topological statistics (*Clustering Coefficient*: 0.864; *Network Diameter*: 18; *Graph Density*: 0.000198; *Average Path Length*: 6.253). Each co-authorship network has only one or two major communities with a few subgroups. We assume that these topological characteristics might be due to the immaturity or the multidisciplinarity of the domain. As a relatively new field, bioinformatics is fast-growing and co-evolving across the sub-communities (Song et al. 2013b). In such a young or highly multidisciplinary domain, it is likely that there is not a coherent field of discipline and a few influential researchers tend to dominate. The number of *Nodes in the Giant Component* also supports this tendency—DBLP: 14,041 nodes which is 51.2 % of the entire nodes; PubMed Central: 49,601 (72.3 %) nodes out of 68,531; WoS: 69,374 (63.9 %). A high *Clustering Coefficient* (DBLP: 0.784; PubMed Central: 0.864; WoS: 0.758) indicates the bioinformatics community is densely connected. Compared to previous measures by Newman (2001) ranging from 0.066 to 0.726, our findings show relatively higher density. The *Average Path Length* also supports the depiction of a close-knit bioinformatics field; researchers whose publications were indexed in PubMed Central are less than seven connections (6.253) apart. In addition, *Network Diameter*, the number of shortest paths between any two nodes in the network, of the PubMed Central network is the shortest—

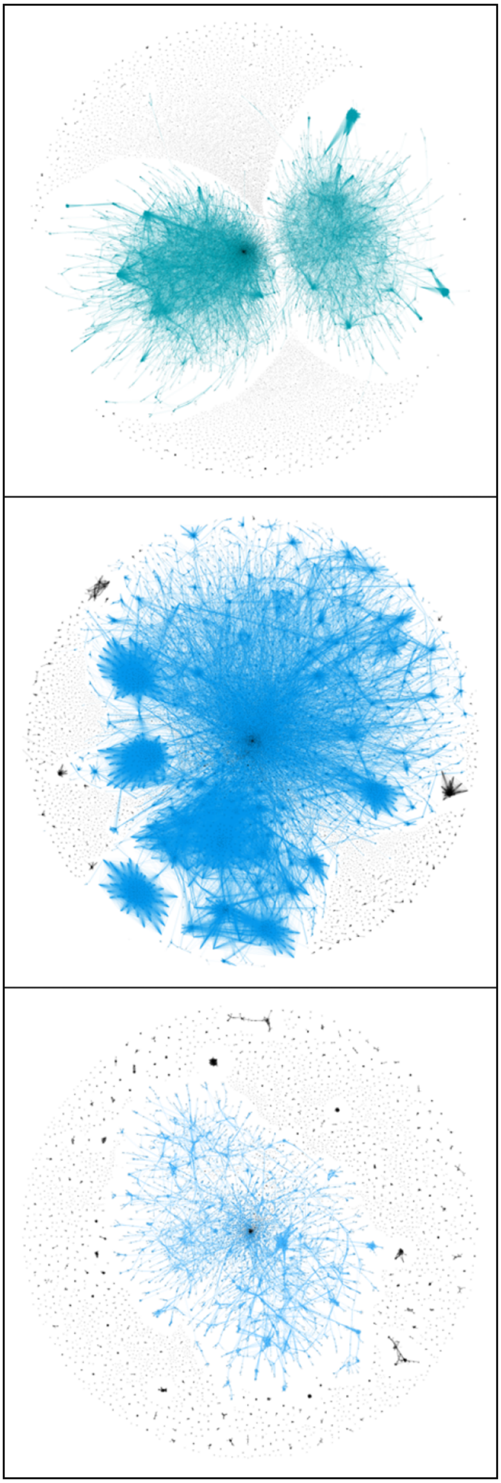

Fig. 5 Co-authorship network of DBLP (*left*), PubMed Central (*center*), and WoS (*right*)

DBLP: 25; PubMed Central: 18; WoS: 27. Overall, these observations illustrate a specific aspect of the landscape of bioinformatics that (1) a few influential researchers tend to lead the academic advancement since the field is emerging and developing, and (2) most of the researchers are intellectually linked within a few steps in spite of the domain's interdisciplinary characteristics; all three data sets show high clustering coefficient, and short average path length and network diameter.

Like Zipf's Law, the relationship between co-authorship frequency and degree centrality obeys a power law distribution (see Fig. 6), the more frequently the researchers collaborate, the more influential they are in the field of bioinformatics. This indicates that a few influential scientists in the field of bioinformatics tend to conduct most of the studies, accompanying peer researchers.

Next, we constructed a complete co-authorship network through single dimensional data integration from all three sources. For data integration, we tagged the names of the databases with co-author pairs, and accumulated the pairs. In addition, we reduced the scale of the network to include only co-author pairs which appeared more than ten times. That is, the threshold for edge weight is 10. Previous studies (Börner et al. 2006; Small 1973; Zizi and Beaudouin-Lafon 1994) employed this straightforward and easy approach to implement link reduction. We compared the entire network with the reduced one which consisted of more influential researchers in terms of edge weight. Some 'small and isolated' clusters were filtered to refine the topological map. Table 4 compares the two types of complete co-authorship networks.

Figures 7 and 8 show the map of the reduced complete co-authorship network. As described in Table 4, each version of the complete co-authorship network demonstrates slightly different characteristics.

First, the complete co-authorship network suggests that more than half (67.5 %) of the researchers in this domain were part of a single large cluster. In addition, the calculation of *Clustering coefficient* of the full complete co-authorship network indicates that the bioinformatics community is a small world (0.655). Compared to previous literature (Newman 2001), the measure of average path length (8.968) is relatively higher. However, any researcher could still reach to all other peers through at the most nine personal connections.

In contrast to the complete co-authorship network, the reduced topology shows a distinguishable pattern: a dichotomous pattern dominated by PubMed Central and WoS (See Fig. 7). In addition, each of the clusters consists of the giant research community. This indicates different research interests lie in both databases. Even in the combined dataset, it is also assumed that bioinformatics is still developing or there is not a coherent body of discipline in bioinformatics. As discussed earlier, previous works observed that this field is co-evolving among sub-/inter-disciplinary fields and also diversifying (Song et al. 2013a, b). If we use a single database such as WoS, we only might reproduce these previously reported trends. We argue, however, different views (e.g. converging, dichotomous, etc.) emerge if we use integrated data sources and identify a more complete landscape. In addition, the network derived from DBLP disappeared behind that of PubMed Central (see Fig. 7). As illustrated in Fig. 8, this pattern may suggest that authors whose manuscripts were indexed by PubMed Central tended to submit their manuscripts to the conferences and journals listed in DBLP. That is, PubMed Central is more inclusive than DBLP. In reality, the number of duplicated authors between PubMed and DBLP were 4,160 whereas 127 between WoS and DBLP in our data collection. There were some additional findings from the reduced network: (1) less than 20 % of the researchers have almost half of the connections (47.9 %), (2) influential co-authors who collaborated more than 10 times consist of a small world (*Clustering coefficient:* 0.78; *Network Diameter*: 22 instead of 28;

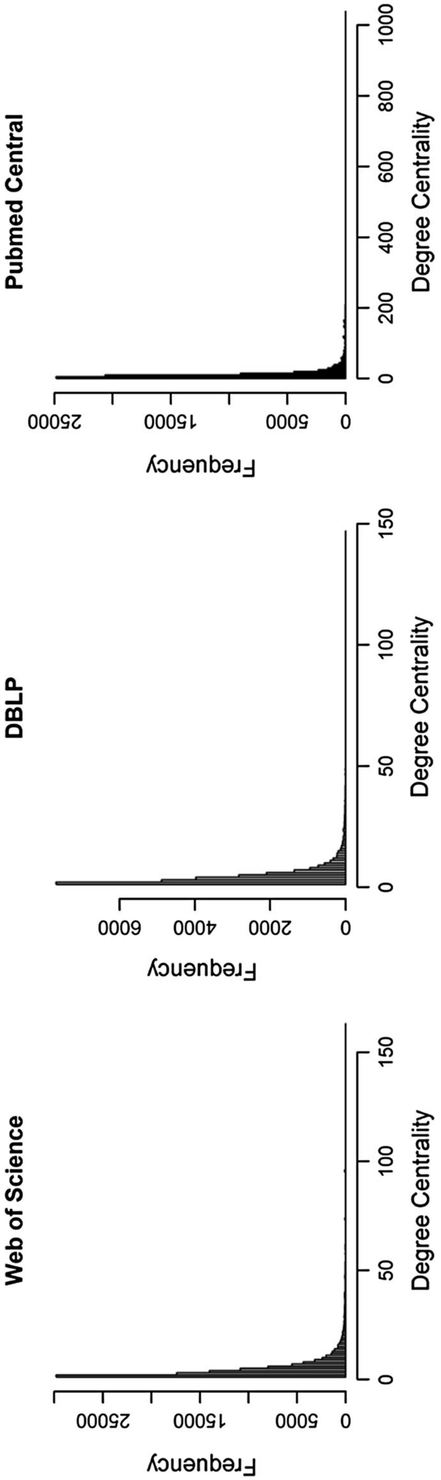

Fig. 6 Power law distribution tendency of co-authorship frequency–degree centrality

Table 4 Complete co-authorship network statistics

	Complete network (full)	Complete network (reduced)
# of nodes	200,043	38,929 (19.5 %)
# of edges	855,803	410,043 (47.9 %)
# of components	14,666	1
Nodes in the giant component	135,109 (67.5 %)	38,929
Clustering coefficient	0.655	0.785
Network diameter	28	22
Graph density	0.000427	0.000541
Average path length	8.968	7.014

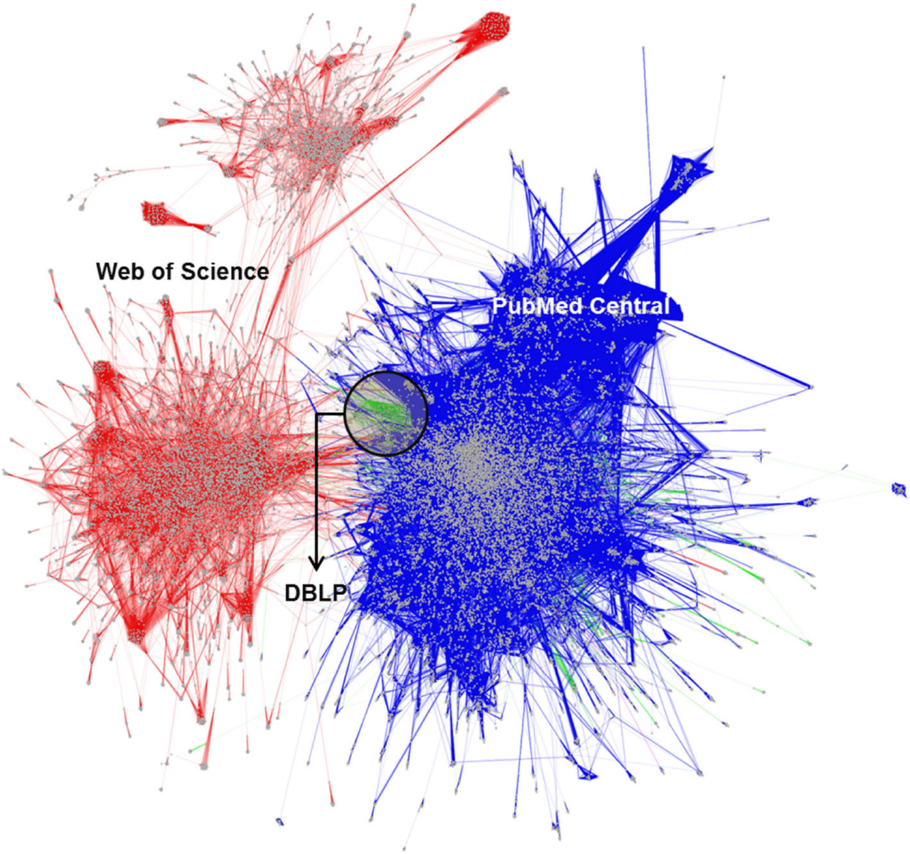

Fig. 7 Complete co-authorship network (reduced)

Average Path Length: 7.014) and are more densely connected (*Graph Density*: 0.000541). Compared to previous literature (Newman 2001), these observations further support the aforementioned finding that most of the researchers in bioinformatics are intellectually

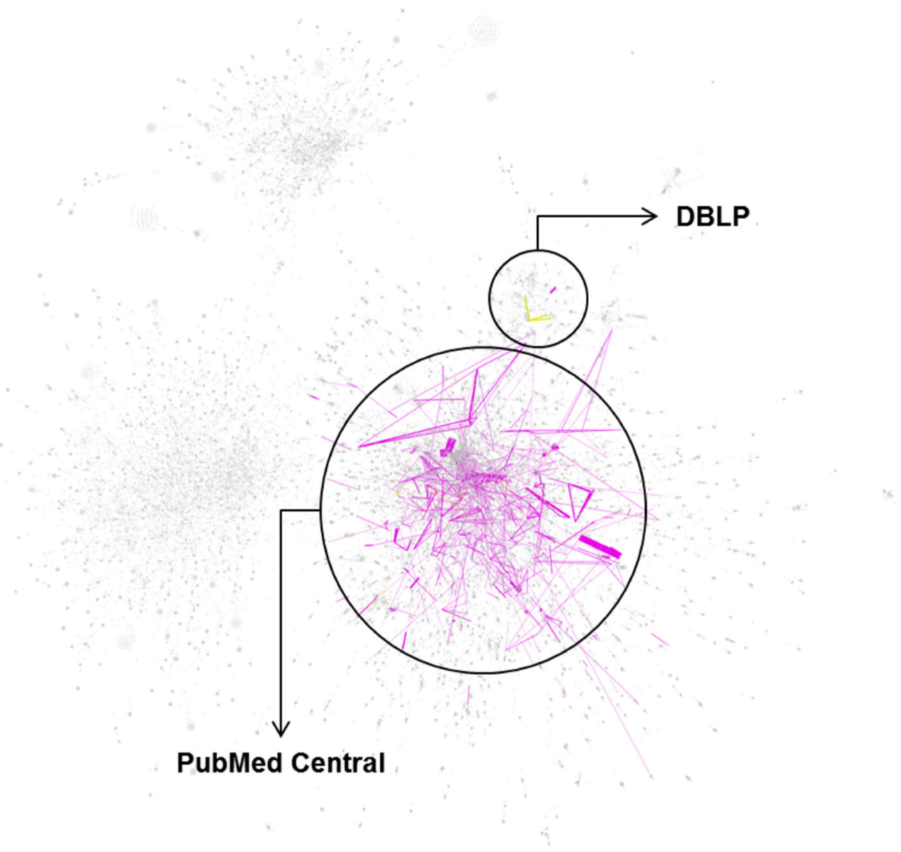

Fig. 8 Highlight on the networks from PubMed Central and DBLP

linked within a relatively few steps in spite of the domain's interdisciplinary characteristics.

The authors who have the highest degree centrality and number of co-authors, like H.-erich Wichmann, Leena Peltonen and Thomas Illig, tend to be located on the cluster of PubMed Central in Fig. 7 whereas such concentration is not observed in highly ranked authors for Web of Science nor DBLP. Table 5 shows the top 10 authors with high degree centrality for each database. There are few overlapping high influential researchers across the databases. This further supports our finding that different academic interests lie in each database, leading to the dichotomous pattern of the scholastic community. It also might be due to the multidisciplinarity of the domain or there is not a coherent discipline of bioinformatics.

The findings from the co-authorship analysis are summarized as follows: (1) the complete landscape shows the dichotomous trend of the domain, (2) a few influential researchers tend to lead the academic advancement, (3) it is likely that most of the researchers know each other in this field, (4) PubMed Central is more inclusive than DBLP as an indexing database, (5) influential co-authors get involved in most of the invisible

Table 5 Top ten influential researchers from each database (co-author network)

Rank	PubMed Central	Web of Science	DBLP
1	Wichmann, H.	Valencia, A.	Thompson, P. M.
2	Peltonen, L.	Smedley, D.	Toga, A. W.
3	Illig, T.	Walker, M.	Rueckert, D.
4	Chanock, S. J.	Kepes, F.	Wong, S. T. C.
5	Van Duijn, C. M.	Norris, V.	De Zubicaray, G. I.
6	Hunter, D. J.	Sato, S.	Ayache, N.
7	Meitinger, T.	Wang, J.	Hawkes, D. J.
8	Poustka, A.	Yuji, N.	Lee, A. D.
9	Wareham, N.	Levy, D.	Frangi, A. F.
10	Hayashizaki, Y.	Marra, M. A.	Madsen, S. K.

communities, and consist of the small world, and (6) different high centrality authors occur in different source clusters.

Co-citation analysis

As described in the above, DBLP does not include any citation information. Therefore, we excluded the DBLP dataset from the co-citation analysis.

In the same manner as the complete co-authorship network, we produced a reduced co-citation network consisting of co-cited author pairs which appeared more than ten times (the threshold for edge weight is ten). We employed this approach because the complete co-citation network was very large; the number of edges with edge weight 1 were 76.9 % (10,159,489), and 85.2 % (22,106,578) of the entire links from PubMed Central (13,209,959), and WoS (25,936,344) respectively. In addition, co-citation pairs with edge weight greater than 10 were selectively removed if they were isolated (not connected with major components). The reduced networks were as follows: the PubMed Central network had 14,424 nodes and 88,534 edges (6.78 and 0.67 % of the entire 212,598 nodes and 13,242,699 edges) and the WoS network had 13,052 nodes and 34,934 edges (3.49 and 0.13 % of the entire 373,952 nodes and 25,936,344 edges). Table 6 summarizes the properties of the co-citation networks.

Reduced to influential pairs whose edge weight is more than 10, these networks show characteristics derived from their source databases, and are distinguishable from each other (see Table 6). In addition, the *Giant Components* include far more nodes than those of the co-authorship networks—PubMed Central: 211,881 nodes which is 99.6 % of the entire nodes; WoS: 373,870 (99.98 %) nodes out of 373,952. This indicates that the majority of scientists in the field of bioinformatics are involved in one huge community. We argue that citers in bioinformatics tend to recognize common citees as their major references. This might be attributed to that bioinformatics is an emerging domain. On the other hand, *Clustering Coefficients* for both full topologies are lower than those of the co-authorship network. However, the co-citation networks are still a small world according to *Network Diameter*s and *Average Path Length*s. These topologies are smaller than the co-authorship topologies (see Tables 3, 4). The co-citation networks are more densely connected than the co-authorship networks (PubMed Central: 0.000499; WoS: 0.000897). In terms of the maximum value of the edge weight, 'Altschul, S. F. & Pearson, W. R.' and 'Altschul, S. F. & Thompson, J. D.' appeared most frequently from each databases (587 and 431 times

Table 6 Co-citation network statistics

	PubMed Central	Web of Science
# of nodes (full)	212,598	373,952
# of nodes (reduced)	14,424 (6.78 %)	13,052 (3.49 %)
# of edges (full)	13,242,699	25,936,334
# of edges with edge weight 1	10,159,489 (76.9 %)	22,106,578 (85.2 %)
# of edges (reduced)	88,534 (0.67 %)	34,934 (0.13 %)
# of components (sub-clusters)	100	27
Nodes in the giant component	211,881 (99.6 %)	373,870 (99.98 %)
Clustering coefficient (reduced)	0.504	0.339
Network diameter (reduced)	17	18
Graph density (reduced)	0.000499	0.000897
Average path length (reduced)	4.203	5.698

DBLP was excluded for this analysis because it does not have citation data

from PubMed Central and WoS). The researcher “Pearson, W. R.” who was found to be most frequently cited, is not the statistician Pearson but a professor of biochemistry and molecular genetics in University of Virginia, School of Medicine. His research paper titled “Improved Tools for Biological Sequence Comparison” has more than 10,000 citations. His most frequently co-cited author, “Altschul, S. F.” is also one of the most influential scientists in this field whose article titled “Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs” was cited more than 40,000 times. We argue this supports our finding that different academic interests lie in each database or the domain is highly multidisciplinary.

The relationship between co-citation frequency and degree centrality follows a power law; the more frequently researchers are co-cited, the more influential they are in the field of bioinformatics (see Fig. 9). As in the co-authorship network, this reveals that a few influential scientists tend to be co-cited by peer researchers.

Figures 10 and 11 show the downscaled map of the co-citation network, labeled with heterogeneous sub-clusters. We assigned labels to the clusters based on the following triangulated procedure: (1) based on the degree centrality, the most influential 10 authors in each cluster were identified, (2) the titles of the articles by those researchers in our data collection and the author’s profiles on their websites were examined, and 3) we manually extracted frequently occurred terms from these two criteria. The labels reveal differences in terms of trends in the studies which were published by the journals and conferences indexed in each database.

The labels in Fig. 10 show that the frequently cited researchers whose citing works were indexed in the PubMed Central database focus on these research topics Based on these labels, major themes in bioinformatics can be classified into two groups: (1) biomedical concepts and entities (Gene Expression, HIV, Infectious Diseases, Metabolism, Neuron, Tumors), and (2) research methodologies and sub-field (Database/Software, Molecular Biology, Ontology/Network, Statistical Analysis). The map does not align with this conceptual categorization, however. For instance, researchers who study Molecular Biology tend to have interests in HIV, Neuron, and Tumors, whereas the counterparts whose research emphasizes Database/Software and Ontology/Network focus on Gene Expression, Infectious Diseases, and Metabolism (see Fig. 10).

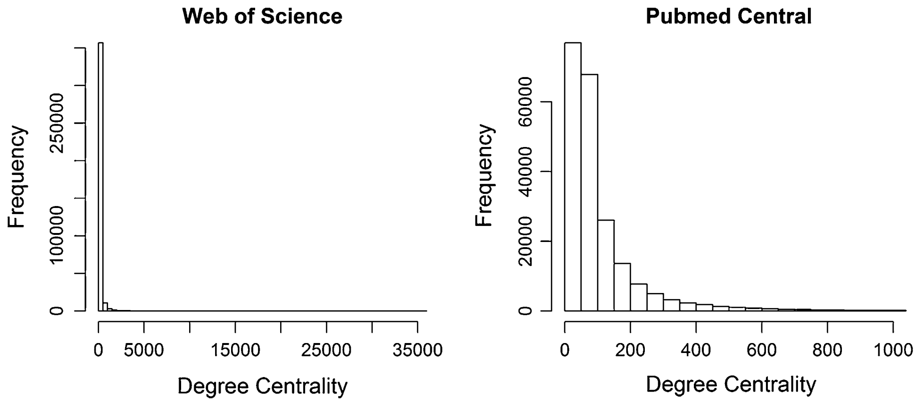

Fig. 9 Power law distribution tendency of co-citation frequency–degree centrality

The co-citation network derived from Web of Science (Fig. 11) differs from the Fig. 10. First, different concepts such as Caenorhabditis Elegans, Chromosome, Cryptococcus, Drosophila, Neurology, Protein, Proteomic, and Statistical Analysis appeared on the map. Except for two methodological concepts, Neurology and Statistical Analysis, the Web of Science network includes biomedical entities which do not appear on the former map. This indicates that the influential scientists whose manuscripts were indexed in the Web of Science database cover different research topics from those on the PubMed Central network. In addition, one of the widely employed approaches in bioinformatics is "Statistical Analysis" (see Fig. 11). On both networks, however, Statistical Analysis is a major subcluster. Based on this, we can assume that statistics is regarded as an essential research methodology in bioinformatics.

Overall, these two dissimilar maps further support our finding that the field of bioinformatics shows topical dichotomy dominated by two different databases, PubMed Central and WoS.

Next, we constructed a complete co-citation network through single dimensional data integration of both sources, and visualized it for further analysis. As mentioned earlier, we tagged the names of the databases with co-citation pairs, and accumulated the pairs for data integration. In addition, we reduced the scale of the network to that with co-author pairs which appeared more than ten times. That is, the threshold for edge weight is also 10 in this case. Some 'small and isolated' clusters were filtered to refine the topological map. Table 7 illustrates the statistics for the complete co-citation network.

Figure 12 above depicts the visualized network of the reduced complete co-citation topology. As reiterated in Table 7, the reduced complete co-citation network indicates that almost every frequently cited researcher (95.83 % of the entire nodes, 17.57 times cited in average) in this domain was included in the largest component. This indicates that there are a few core influential researchers who tend to be cited by peers due to the fact that the field is still emerging. It seems true in every science and we also identify that the similar trend lies in the field of bioinformatics. Table 7 also shows that the *Clustering Coefficient* of the complete co-citation network (0.146) is lower than that of the complete co-authorship network (0.785). Based on this, we might argue that after integrating major data sources in bioinformatics, a small world phenomenon is not very vividly observed. That is, it is likely

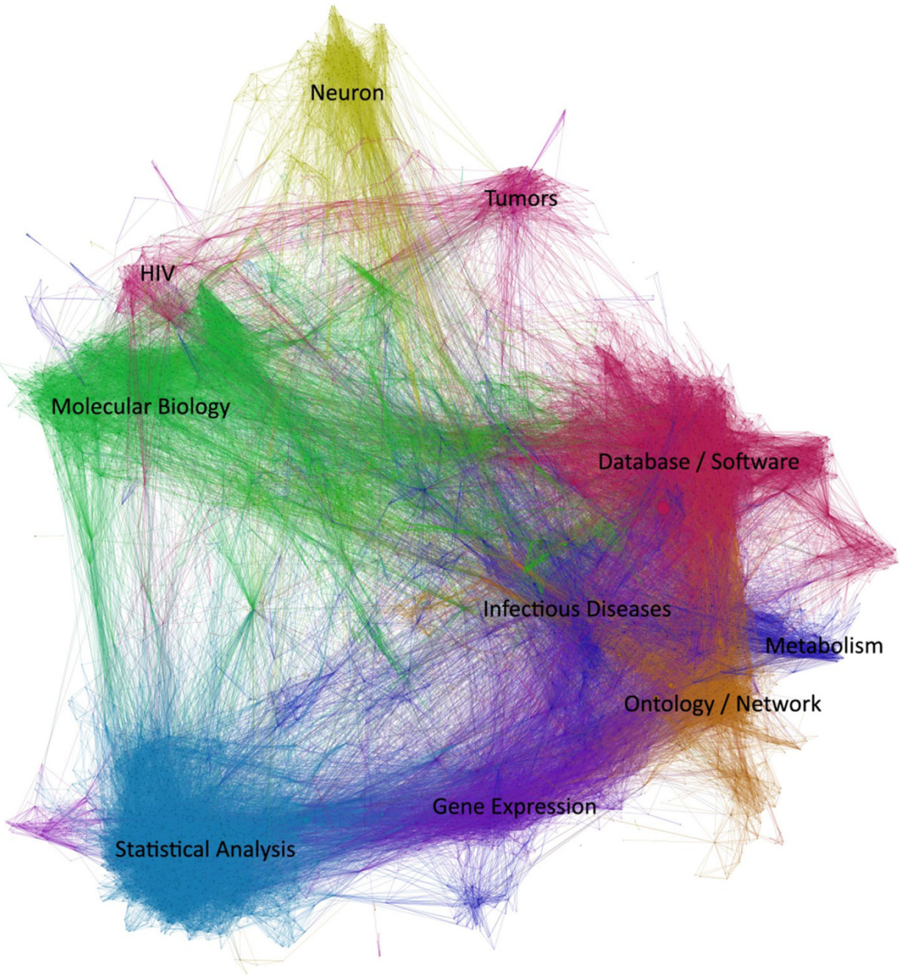

Fig. 10 Co-citation topology of PubMed Central (reduced)

that collaborative authors tend to have more intellectual ties whereas co-cited researchers might be referred by indirect purposes. However, the *Average Path Length* demonstrates that any researcher could reach to other peers in the network within five personal connections (*Average Path Length*: 4.823). That is, authors are closely connected but retain their most direct ties to their respective disciplinary areas.

Interestingly, this reduced topology also shows a dichotomous pattern; this map is clearly divided into two sections representing the two databases, but forms a more complete integrated picture when combined. This supports the interesting finding from the complete co-authorship network analysis that two databases, PubMed Central and Web of Science, cover different aspects of bioinformatics, allowing few overlaps (see also Fig. 7). Together with the findings from the co-citation analysis, this finding further emphasizes that each database covers different research topics (see Figs. 10, 11). In addition, these

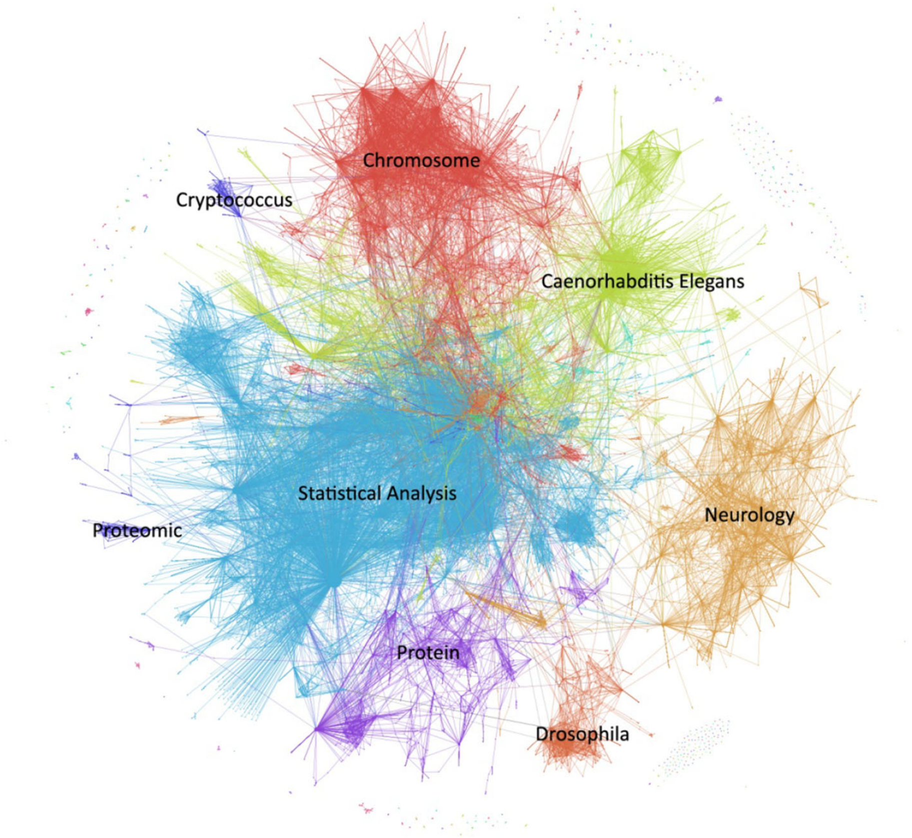

Fig. 11 Co-citation topology of Web of Science (reduced)

Table 7 Complete co-citation network statistics (reduced)

	Complete network (reduced)
# of nodes	31,940 (5.87 %)
# of edges	164,494 (0.4 %)
# of components	403
Average edge weight	17.57
Average degree centrality	9.932
Nodes in the giant component	30,608 (95.83 %)
Clustering coefficient	0.146
Network diameter	17
Graph density	0.000322
Average path length	4.823

findings suggest that in the bibliometric study of bioinformatics, both databases need to be considered in conjunction with one another. Overall, this observational result implies that each data source reveals only part of the whole picture of bioinformatics.

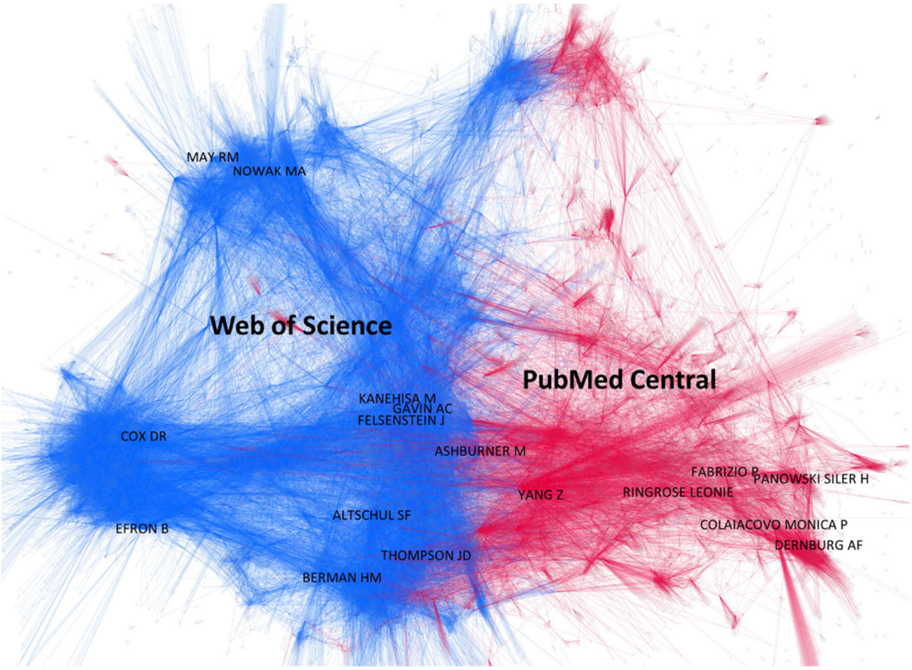

Fig. 12 Complete co-citation network (reduced)

Table 8 Top 10 influential researchers from each database (co-citation network)

Rank	PubMed Central	Web of Science
1	**Altschul, S. F.**	**Altschul S. F.**
2	Efron, B.	Colaiacovo, M. P.
3	Cox, D. R.	**Thompson, J. D.**
4	**Ashburner, M.**	Dernburg, A. F.
5	Berman, H. M.	Panowski, S. H.
6	Kanehisa, M.	Ringrose, L.
7	Felsenstein, J.	Yang, Z.
8	May, R. M.	**Ashburner, M.**
9	Nowak, M. A.	Anne-Claude, G.
10	**Thompson, J. D.**	Paola, F.

Overlapped authors between databases are bolded

Figure 12 shows top 10 influential authors in each database. The authors highly co-cited in both databases (Altschul, S. F., Ashburner, M., and Thompson, J. D.) lie in center of this network. Table 8 shows the list of those authors with high degree centrality in each database; noticeably, there are few overlaps. Some of the most highly influential authors do overlap between both databases however, perhaps because their studies have a more generally applicable impact. This finding largely supports the aforementioned dichotomous research trend.

Overall, the findings from the co-citation analysis are similar those from the co-authorship analysis. First, the complete landscape shows the dichotomous pattern with distinct research interests covered by each database. Second, a few influential scientists in the field of bioinformatics tend to be co-cited by peer researchers because the domain is an emerging field. Third, it is likely that collaborative authors tend to have more intellectual ties whereas co-cited researchers might be referred by indirect purposes. Finally, the bioinformatics community is a small world when compared to previous research.

Discussion and conclusion

We conducted bibliometric analysis to explore how various data sources with different scientific objectives influence the formation of invisible communities. We also tried to reveal a more complete landscape of bioinformatics by integrating the bibliometric networks derived from these different sources. We specifically focused on the domain of bioinformatics and employed two different types of bibliometric analysis: (1) co-authorship analysis, and (2) co-citation analysis. The major contribution of the study is discovering how various topological characteristics relate to concepts of community that we observed in the present study.

Our study also identified some interesting findings on the author network in the field of bioinformatics.

First, the complete landscape of bioinformatics from the integrated data sources illustrates a dichotomous pattern dominated by PubMed Central and WoS. This result implies that the selection of a particular database (e.g. WoS) will lead to a partial interpretation of the academic landscape while highlighting specific aspects. It also supports our finding that there is not a coherent body of knowledge in bioinformatics. Second, a few influential scientists in the field of bioinformatics receive very high citations from their colleagues, which is a driving force to bloom the field. In particular, those influencers are most likely to conduct studies accompanying peer researchers. Third, the majority of scientists in the domain are involved in a few giant research community. This finding would be a commonly recognized phenomenon in any intellectual or social community. However, we empirically confirmed this feature of the field of bioinformatics in terms of the complete landscape produced from the dataset synthesized from multiple sources. We believe that the main contribution of the study is more about the methodology of exploring big citation data sets. Thus, employing our approach for work on other emerging and developing fields is beneficial for mapping more complete landscapes. Fourth, influential authors form a smaller world compared to other researchers. Fifth, the journals and conferences indexed by each database cover different research topics, and PubMed Central is more inclusive than DBLP as an indexing database. Lastly, our research findings suggest that both PubMed Central and Web of Science should be used together for any future bibliometric studies of bioinformatics. These findings need to be confirmed by conducting bibliometric analysis of different biomedical domains indexed in both databases.

It must be noted note that our approach to identifying unique authors from the DBLP and Web of Science datasets would not discriminate between researchers whose names share the same abbreviation. This problem arises from how DBLP and Web of Science organize their bibliographic information, and was not able to be remedied through our methodology. Although bibliographic descriptions supported by both databases do not include additional identifiers in the author field, this problem may be able to be addressed if additional descriptors from the full texts are utilized; our future research will investigate the feasibility of this method. Despite the intuitiveness of our approach to label the heterogeneous clusters, automated cluster labeling may produce more representative terms by drawing on the entire dataset. We also hope to tackle this challenge in a future study.

In the future, we intend to investigate how the bibliometric patterns we have identified vary over time. We will also explore whether the community structure detected is statistically correlated with the results of the topological measurements reported in this paper. We hope that a similar analysis with a second corpus and generalize the result of the present study beyond the landscape of bioinformatics.

Acknowledgments This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Ministry of Science, ICT & Future Planning (Grant No. 2013M3A9C4078138).

References

Abraham, A., Hassanien, A., & Snasel, V. (2010). *Computational social network analysis: Trends, tools and research advances*. New York: Springer.

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *JASIST, 54*(6), 550–560.

Barnett, G. A. (Ed.). (2011). *Encyclopedia of social networks*. Thousand Oaks, CA: Sage.

Beaver, D., & Rosen, R. (1978). Studies in scientific collaboration. Part I. The professional origins of scientific co-authorship. *Scientometrics, 1*, 65–84.

Beaver, D., & Rosen, R. (1979). Studies in scientific collaboration. Part II. Scientific co-authorship, research productivity and visibility in the French Elite. *Scientometrics, 1*, 133–149.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 10*, 10008.

Börner, K., Penumarthy, S., Meiss, M., & Ke, W. (2006). Mapping the diffusion of scholarly knowledge among major U.S. research institutions. *Scientometrics, 68*(3), 415–426.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117.

Catala-Lopez, F., et al. (2012). Coauthorship and institutional collaborations on cost-effectiveness analyses: A systematic network analysis. *PLoS One, 7*(5), e38012.

Chen, F., Chen, Z., Wang, X., & Yuan, Z. (2008). The average path length of scale free networks. *Communications in Nonlinear Science and Numerical Simulation, 13*(7), 1405–1410.

Chung, F. R. K. (1984). Diameters of communications networks. Mathematics of Information Processing, AMS Short Course Lecture Notes, 1–18.

Clarke, B. L. (1964). Multiple authorship trends in scientific papers. *Science, 143*, 822–824.

Clarke, B. L. (1967). Communication patterns of biomedical scientists. *Federation Proceedings, 26*, 1288–1292.

Coleman, T. F., & Moré, J. J. (1983). Estimation of Sparse Jacobian matrices and graph coloring problems. *SIAM Journal on Numerical Analysis, 20*(1), 187–209.

Day, M., Ong, C., & Hsu, W. (2010). An analysis of research on information reuse and integration (2003–2008). *ITSSA, 6*(2), 146–157.

Ding, Y., Yan, E., Frazho, A., & Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. *JASIST, 60*(11), 2229–2243.

Erma, N., & Todorovski, L. (2010). Co-authorship network analysis in the E-government research field. In *Proceedings of EGOV '10*.

Glänzel, W., Janssens, F., & Thijs, B. (2009). A comparative analysis of publication activity and citation impact based on the core literature in bioinformatics. *Scientometrics, 79*(1), 109–129.

Glänzel, W., & Schubert, A. (2004). Analyzing scientific networks through co-authorship. Handbook of Quantitative Science and Technology Research, 257–276.

Hany, Y., Zhouz, B., Peiz, J., & Jiay, Y. (2009). Understanding importance of collaborations in co-authorship networks: A supportiveness analysis approach. In *Proceedings of SDM '09*.

He, B., Tang, J., Ding, Y., Wang, H., Sun, Y., et al. (2011). Mining relational paths in integrated biomedical data. *PLoS One, 6*(12), e27506.

Heffner, A. G. (1981). Funded research, multiple authorship, and subauthorship collaboration in four disciplines. *Scientometrics, 3*, 5–12.

Hou, H., Kretschmer, H., & Liu, Z. (2006). The structure of scientific collaboration networks in scientometrics. In *Proceedings of COLLECT'06*.

Huang, H., Andrews, J., & Tang, J. (2012). Citation characterization and impact normalization in bioinformatics journals. *JASIST, 63*(3), 490–497.

Huang, T., & Huang, M. L. (2006). Analysis and visualization of co-authorship networks for understanding academic collaboration and knowledge domain of individual researchers. In *Proceedings of IEEE CGIV '06* (pp. 18–23).

Ioannidis, J. P. A. (2008). Measuring co-authorship and networking-adjusted scientific impact. *PLoS One, 3*(7), e2778.

Janssens, F., Glänzel, W., & De Moor, B. (2007). Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. In *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 360–369).

Kim, H. & Barnett, G. A. (2008). Social network analysis using author co-citation data. In *Proceedings of the 14th AMCIS*, 172.

Kolchinsky, A., Abi-Haidar, A., Kaur, J., Hamed, A. A., & Rocha, L. M. (2010). Classification of protein–protein interaction full-text documents using text and citation network features. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7*(3), 400–411.

Kulkarni, A. V., Aziz, B., Shams, I., & Busse, J. W. (2009). Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. *JAMA, 302*(10), 1092–1096.

Kumar, R., Novak, J., & Tomkins, A. (2010). *Structure and evolution of online social networks*. NY: Springer.

Leydesdorff, L. (2005). Similarity measures, author co-citation analysis, and information theory. *JASIST, 57*(7), 769–772.

Leydesdorff, L., & Vaughan, L. (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the web environment. *JASIST, 57*(12), 1616–1627.

Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods of Information in Medicine, 40*(4), 346–358.

Luukkonen, T., Persson, O., & Silvertsen, G. (1992). Understanding patterns of international scientific collaboration. *Science, Technology and Human Values, 17*, 101–126.

Luukkonen, T., Tijssen, R. J. W., Persson, O., & Silvertsen, G. (1993). The measurement of international scientific collaboration. *Scientometrics, 28*, 15–36.

Manoharan, A., Kanagavel, B., Muthuchidambaram, A., & Kumaravel, J. P. S. (2011). Bioinformatics research—An informetric view. *International Conference on Information Communication and Management, 2011*, 199–204.

Marques-Pita, M., & Rocha, L. M. (2013). Canalization and control in automata networks: Body segmentation in Drosophila melanogaster. *PLoS One, 8*(3), e55946.

Morel, C. M., Serruya, S. J., Penna, G. O., & Guimaraes, R. (2009). Co-authorship network analysis: A powerful tool for strategic planning of research, development and capacity building programs on neglected diseases. *PLoS, 3*(8), e501.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *PNAS, 98*(2), 404–409.

Newman, M. E. J. (2004). Co-authorship networks and patterns of scientific collaboration. *PNAS, 101*, 5200–5205.

Nooy, D. W., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. New York: Cambridge University Press.

Patra, D., Mishra, A. K. (2006). Chemical and biochemical fluorescence sensors encyclopedia of sensors. In C. A. Grimes, E. C. Dickey & M. V. Pishko (Eds.) (Vol. 2, pp. 139–156). CA, USA: American Scientific Publishers.

Perez-Iratxeta, C., Andrade-Navarro, M. A., & Wren, J. D. (2007). Evolving research trends in bioinformatics. *Brief Bioinform, 8*(2), 88–95.

Perry, C. A., & Rice, R. E. (1998). Scholarly communication in developmental dyslexia: Influence of network structure on change in a hybrid problem Area. *JASIST, 49*, 151–168.

Persson, O. (2001). All author citations vs first author citations. *Scientometrics, 50*, 339–344.

Price, D., & Beaver, D. (1966). Collaboration in an invisible college. *American Psychologist, 21*, 1011–1018.

Rousseau, R., & Zuccala, A. (2004). A classification of author cocitations: Definitions and search strategies. *JASIST, 55*, 513–529.

Sade, D. S. (1989). Sociometrics of Macaca Mulatta III: N-path centrality in grooming networks. *Social Networks, 11*, 273–292.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between publications. *JASIST, 24*, 265–269.

Smith, M. (1958). The trend toward multiple authorship in psychology. *American Psychologist, 13*, 596–599.

Song, M., Kim, S. Y., Zhang, G., Ding, Y., & Chamber, T. (2013a). Productivity and influence in bioinformatics: A bibliometric analysis using PubMed Central. *JASIST, 65*(2), 352–371.

Song, M., Yang, C. C., & Tang, X. (2013b). Detecting evolution of bioinformatics with a content and co-authorship analysis. *Springerplus, 2*(1), 186.

The DBLP Website. Retrieved August 18, 2013 from http://www.informatik.uni-trier.de/~ley/db/.

The PubMed Central Website. Retrieved August 18, 2013 from http://www.ncbi.nlm.nih.gov/pmc/.

The Web of Science. Retrieved August 18, 2013 from http://thomsonreuters.com/web-of-science/.

Velden, T., Haque, A., & Lagoze, C. (2009). A new approach to analyzing patterns of collaboration in co-authorship networks—Mesoscopic analysis and interpretation. In *Proceedings of ISSI '09* (pp. 14–17).

Wasserman, S., & Katherine, F. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

Watts, D. J., & Strogatz, S. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440–442.

White, H. D. (2003a). Author cocitation analysis and Pearson's R. *JASIST, 54*(13), 1250–1259.

White, H. D. (2003b). Pathfinder network and author cocitation analysis: A remapping of paradigmatic information scientists. *JASIST, 54*(5), 423–434.

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *JASIST, 32*, 163–172.

White, H. D., Wellman, B., & Nazer, N. (2004). Does citation reflect social structure? *JASIST, 55*(2), 111–126.

Zhao, D. (2006). Towards all-author co-citation analysis. *Information Processing and Management, 42*, 1578–1591.

Zizi, M., & Beaudouin-Lafon, M. (1994). Accessing hyperdocuments through interactive dynamic map. In *Proceedings of the 1994 ACM European conference on Hypermedia technology* (pp. 126–135).