Comparative Evaluation of Bibliometric Content Networks by Tomographic Content Analysis: An Application to Parkinson's Disease

Keeheon Lee
Creative Technology Management, Underwood International College, Yonsei University, Seoul, Korea.
E-mail: keeheon@yonsei.ac.kr

SuYeon Kim
Department of Library and Information Science, Kyonggi University, 154-42, Gwanggyosan-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, Republic of Korea.
E-mail: suyeon@yonsei.ac.kr

Erin Hea-Jin Kim and Min Song
Department of Library and Information Science, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, Korea.
E-mail: erin.hj.kim@yonsei.ac.kr, min.song@yonsei.ac.kr

To understand the current state of a discipline and to discover new knowledge of a certain theme, one builds bibliometric content networks based on the present knowledge entities. However, such networks can vary according to the collection of data sets relevant to the theme by querying knowledge entities. In this study we classify three different bibliometric content networks. The primary bibliometric network is based on knowledge entities relevant to a keyword of the theme, the secondary network is based on entities associated with the lower concepts of the keyword, and the tertiary network is based on entities influenced by the theme. To explore the content and properties of these networks, we propose a tomographic content analysis that takes a slice-and-dice approach to analyzing the networks. Our findings indicate that the primary network is best suited to understanding the current knowledge on a certain topic, whereas the secondary network is good at discovering new knowledge across fields associated with the topic, and the tertiary network is appropriate for outlining the current knowledge of the topic and relevant studies.

This article was published online on 21 December 2016. An error was subsequently identified. This notice is included in the online and print versions to indicate that both have been corrected on 9 March 2017.

Received December 3, 2015; revised March 14, 2016; accepted April 21, 2016

 DOI: 10.1002/asi.23752

Introduction

Both new scientific collaboration and knowledge discovery have gained recent attention as subject matter for bibliometric network analyses. In particular, we focus on bibliometric content networks. Various bibliometric network analyses provide insight into the landscape of a particular topic from multiple angles. Those analyses consist of collecting data sets relevant to the topic by querying knowledge entities (Ding et al., 2013), extracting knowledge entities from the data sets, building a network based on the relationship between the entities, and analyzing the network.

We classify three different networks by their construct data set. First, primary bibliometric network is based on knowledge entities from documents retrieved by a simple keyword. It builds a fundamental knowledge structure of explicit and direct entities to a concept. Most previous studies focus on this network (Chen, Wan, Jiang, & Cheng, 2014; Ho et al., 2013; Li, Ho, & Li, 2008; Ponce & Lozano, 2011; Rodrigues, Fonseca, & Chaimovich, 2000). However, two more extensions can be made by using entities from documents selected based on ontological concepts to the keyword and those from documents cited by the first documents. The former extension is a secondary bibliometric network that consists of direct but implicit entities to the concept. The entities often mediate the concept with other concepts. The latter extension is a tertiary bibliometric network that comprises indirect entities to the concept. The concept of entitymetrics, that is, pausing a particular interest

in studying the bibliometric network at the bio entity level, started to expand bibliometrics into the third bibliometric network (Song, Han, Kim, Ding, & Chambers, 2013).

Our goal is to identify similarities and differences among three networks. To achieve this goal, we address the statistical and network properties of the networks, and apply a novel tomographic content analysis. The new analysis allows us to dissect network content by topological layers, whereas the conventional content analysis focuses on central terms relevant to a certain theme only. The layer consists of two axes representing local importance (i.e., degree) and global importance (i.e., betweenness centrality). The range of each layer is determined by applying Zipf's law to grouping the entities. Each layer is divided into four parts for descriptive analysis at the entity level. In this approach, we measure how similar the networks' layers are by the similarity of their entity vectors. To this end, we compiled Parkinson's disease (PD) data sets from PubMed and PubMed Central as a case study. We used MeSH terms as entities and their co-occurrences to build the three networks.

The present article is organized as follows: Related Work outlines the literature review; Methodology details our method; Results evaluates the research results; then the Discussion discusses additional issues; and the final section gives our conclusion and identifies possible future work.

Related Work

Traditional Bibliometric Network Construction

A bibliometric network is widely applied to evaluate scientific output and research performance. It can identify a knowledge domain (Al-Kindi, Al-Juhaishi, Haddad, Taheri, & Khalil, 2015; Ho et al., 2013) and global network of knowledge (Cheng & Zhang, 2013) or assess the scholarly impact of entities such as authors (Ponce & Lozano, 2011), institutions (Chen et al., 2014), and countries (Bramness, Henriksen, Person, & Mann, 2013; Rodrigues et al., 2000), or we can understand historical perspective through it (Li et al., 2008). These approaches analyze research trends (Blockmans, Engwall, & Weaire, 2014; Lewin, 2008) and/or individual, institutional, or national productivity (Li et al., 2012; Royle & Waugh, 2015; Uthman et al., 2015; Yang et al., 2013; Zyoud, Al-Jabi, Sweileh, & Awang, 2014). Sometimes it deals with funding assessment (Lewison, Rippon, De Francisco, & Lipworth, 2004; Van Leeuwen, Van der Wurff, & Van Raan, 2001; Yang et al., 2013). The Web of Science database (WoS) is the most widely used data set (Gu et al., 2015; Schäfer, Hiemke, & Baumann, 2015), whereas the SCOPUS database (Cheng & Zhang, 2013) or PubMed database (Ugolini et al., 2013) takes up a small fraction of the source of data sets. Knowledge entities, such as keywords in the title or abstract (Ho et al., 2013), or provided by authors (Li et al., 2008) or database provider (Cantos-Mateos, Vargas-Quesada, Chinchilla-Rodríguez, & Zulueta, 2012), are used to observe research trends. In biomedical bibliometrics, MEDLINE searches are conducted to use the intended subject headings (Sorensen, Seary, & Riopelle, 2010; Ugolini, Puntoni, Perera, Schulte, & Bonassi, 2007).

Biobibliometric analysis has been performed in various medical fields, such as stem cell (Cantos-Mateos et al., 2012; Ho, Chiu, Tseng, & Chiu, 2003), spine (Ding, Jia, & Liu, 2016), cardiovascular disease (Al-Kindi et al., 2015; Huffman et al., 2013; Ugolini et al., 2013), chronic obstructive pulmonary disease (COPD; Gu et al., 2015), rheumatism (Cheng & Zhang, 2013), diabetes (Krishnamoorthy, Ramakrishnan, & Devi, 2009; Sweileh, Sa'ed, Al-Jabi, & Sawalha, 2014), Severe Acute Respiratory Syndrome (SARS; Rodrigues et al., 2000), small cell lung cancer (SCLC; Ho et al., 2013), and Alzheimer's disease (AD; Chen et al., 2014). Li et al. (2008) explained exponentially growing PD trends, between 1991 and 2006, based on the WoS database, including productivity and international collaboration by analyzing citations and author-provided keywords. The relevant studies are the identification of the most productive PD research over time (Ponce & Lozano, 2011), the highest H-indexed scientists since 1985 (Sorensen & Weedon, 2011), the impact of PD as a neurological disabling disease on rehabilitation (Tesio, Gamba, Capelli, & Franchignoni, 1995), and the productivity and the impact of Indian PD research (Gupta & Bala, 2013). These studies attempted to describe the patterns of research trends in PD by the quantitative analyses of major entities. In our study, we investigate PD profoundly as well as statistically so as to understand PD.

Heterogeneous Entity Network Construction

The traditional biomedical bibliometrics consider common entities (author, journal, or country) in the scientific literature, whereas the literature-based discovery (LBD) of biomedical knowledge extends its capability into domain-level entities (gene, drug, or disease) by text mining. Domain-level entities can be expressed as two types of heterogeneous entity networks: explicit and implicit. An explicit heterogeneous entity network refers to the entity–entity network based on entity co-occurrences in the same publication (Hoffmann & Valencia, 2004; Huang, 2013; Jenssen, Lægreid, Komorowski, & Hovig, 2001; Natarajan et al., 2006; Stapley & Benoit, 2000). An implicit heterogeneous entity network refers to the entity–citation entity network associated with three different entities in two independent publications (Ding et al., 2013; Lee, Kim, Charidimou, & Song, 2015; Song et al., 2013; Song, Heo, & Lee, 2015; Yu et al., 2015).

Explicit heterogeneous entity networks are used for generating a gene/protein network including phenotypes, pathologies, and gene function (Hoffmann & Valencia, 2004), a gene relationship network based on pairwise interaction patterns in molecular biology and biomedicine (Natarajan et al., 2006), and biomarker networks and disease–gene networks of neurological diseases (Huang, 2013), from the PubMed literature. Recently, implicit heterogeneous entity networks are utilized for inferring knowledge entities (i.e., Entitymetrics). Ding et al. (2013) introduced entitymetrics to analyze the entity–entity citation network derived from Swanson's model

DOI: 10.1002/asi

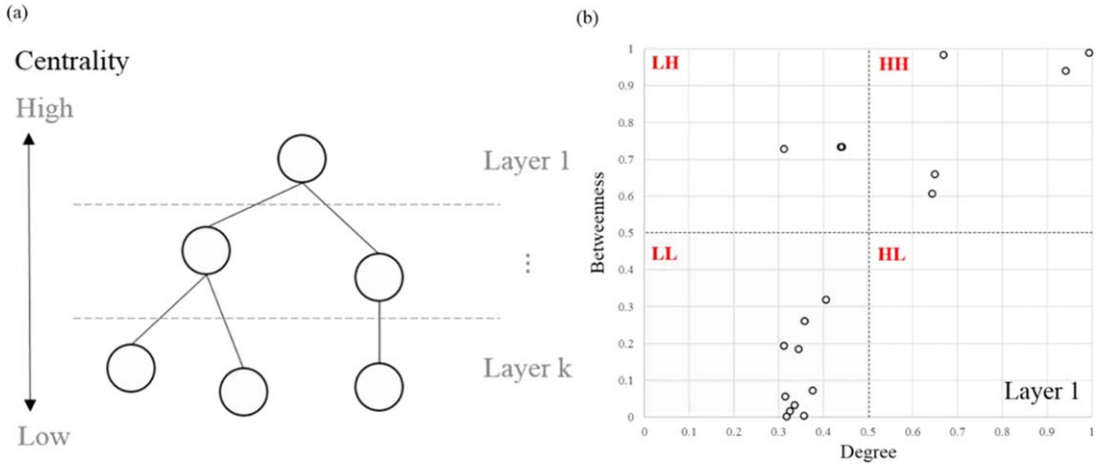

FIG. 1. (a) Layers of a topology based on centrality; (b) example of layer 1 with degree and betweenness centralities. [Color figure can be viewed at wileyonlinelibrary.com]

(Swanson, 1986) based on the metformin-related literature. The practicality of entitymetrics has been investigated by comparing an implicit heterogeneous entity network with the corresponding explicit heterogeneous entity network of gene interactions (Song et al., 2013) and generating various entitymetric networks of Alzheimer's disease (Lee et al., 2015). From the studies, gene–citation–gene (GCG) networks hit 96% coverage of BioGRID (Song et al., 2013) and are shown to be more dense and complex than entity–entity (GG) networks. For PD, Kostoff (2014) attempted to discover potentially linked entities between two different papers through a bibliographic coupling-based network.

In the present study we investigated the effects of querying knowledge entities on the construction of heterogeneous entity networks in the PD literature by our proposed tomographic content analysis. We examine and compare three types of heterogeneous entity networks derived by the explicit and implicit relationship between domain knowledge entities.

Methodology

Three Different Bibliometric Content Networks

We introduce three different bibliometric content networks. Their nodes are knowledge entities, such as keywords (e.g., MeSH terms in MEDLINE), extracted from documents gathered from bibliographic databases, such as WoS or PubMed. Their edges are the frequency of the co-occurrences of two entities in a document. In the case of PD, a primary bibliometric content network (PBN) comprises co-occurring keywords based on publications obtained by a query with terms directly and explicitly associated with a focal topic, PD. This shows the knowledge structure of PD and is utilized in general. However, two more networks can be considered. The secondary bibliometric content network (SBN) consists of those keywords that emerge concurrently based on the publications collected by associated concepts (e.g., proteins) directly but implicitly connected with PD. It illustrates the associations between PD and other diseases mediated by the proteins. Tertiary bibliometric content network (TBN) contains those keywords that appear coincidently in relation to citing and cited publications. TBN depicts knowledge entities indirectly connected with PD.

Tomographic Content Analysis

To compare the three types of bibliometric content networks, we introduce tomographic content analysis. It is a descriptive analysis that essentially dissects the cross-sections of a network. Within the cross-sections, we identify the position of a node in relation to others, the meaning of the position, and the topics made of node clusters (i.e., modules). We also introduce a metric that summarizes the similarity between two layers or two modules.

Positioning. We decompose a network into topological layers to identify the position of a knowledge entity on a layer as in Figure 1a. A layer is two-dimensional plane of degree centrality and betweenness centrality but the ranges of the two centralities are bounded. A layer has larger numbers for the boundaries of the ranges than the lower layer. The ranges are given by dividing the range of the minimum and maximum values of each centrality by two. Assume that the entire range is [0, 6,000]. The first layer is [3,000, 6,000], and the second layer is in the range [1,500, 3,000]. This is because a terminology network has the property of "Zipf's law," which states that the most frequently occurring term has twice the number of the next frequent term. The degree centrality of a node refers to the local importance of the node among neighboring nodes. The betweenness centrality of a node indicates the globally mediating importance of the node among non-neighboring nodes. The position of a node on a layer shows how the node is related to others locally and globally in a certain research topic.

Each layer contains four quadrants. As a result, layer 1 (the white area) in Figure 1b has entities with high degree or betweenness centralities. In each layer, a dot on the upper-

DOI: 10.1002/asi

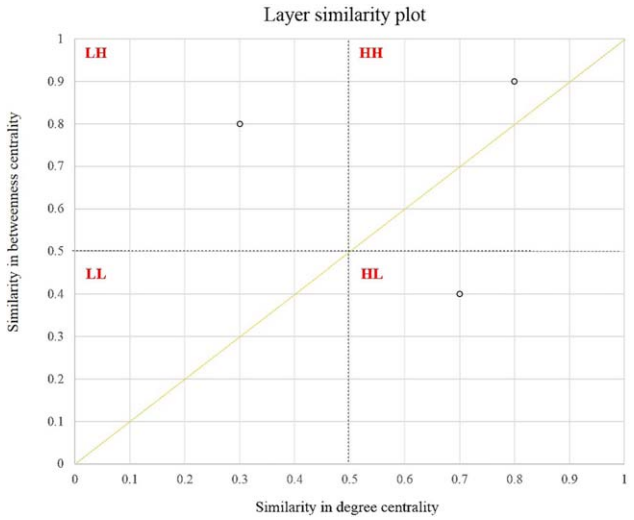

FIG. 2. Layer similarity plot. [Color figure can be viewed at wileyonlinelibrary.com]

right side (i.e., HH quadrant: high degree and high betweenness centrality) represents an impactful entity locally and globally. A dot on the lower-right side (i.e., HL quadrant: high degree and low betweenness centrality) means a locally influential entity. A dot on the upper-left side (i.e., LH quadrant: low degree and high betweenness centrality) implies a globally powerful entity. The dots on the lower-left side (i.e., LL quadrant: low degree and low betweenness centrality) are locally and globally ineffective entities in this layer, and become the entities for the lower layers. Namely, a lower layer is essentially LL part of the previous layer. For example, layer 2 is LL part of Layer 1. Recursively, layer 2 is divided into four quadrants and LL part of layer 2 becomes layer 3. And thus, the last layer covers all the left nodes.

We name a node cluster as a topic. Topics are detected by the community detection algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). The algorithm works as agglomerative clustering. It starts with each node as a cluster, merges clusters if the density of the merged cluster gets higher than before, and iterates merging until no new cluster is made. The topics help conceptualize the member nodes for better understanding. A cluster may be located across layers. A major module tends to be in a higher layer, whereas a minor module is likely to be in a lower layer. The nodes in each module are represented as a unique shape

Comparison. We calculate the content similarity between two layers or two modules for a simple comparison. We select two layers from different networks. A degree vector and a betweenness vector of terms are made for each layer. We conduct cosine similarity between two numerical vectors of two layers of different networks. For each layer comparison, we get two similarity values for two centralities.

Figure 2 depicts the similarity in degree centrality and determines the location with respect to the x-axis. That regarding betweenness centrality decides the location with respect to the y-axis. Similar to Figure 1b, the plot is divided into four parts: LL for low degree similarity and low betweenness similarity, LH for low degree similarity and high betweenness similarity, HL for high degree similarity and low betweenness similarity, HH for high degree similarity and high betweenness similarity. The distance from the origin represents the total similarity between two layers regarding two centralities. A point on HH implies two layers are highly similar in both degree and betweenness centralities, with similar terms. A point on HL means two layers are highly similar in the degree centrality only. In other words, the entities in the two layers serve similar roles locally, but their mediating roles are different globally. A point on LH indicates two layers have similar mediating entities but different phases locally. A point on LL means two layers are different.

When all points fall in a diagonal line, the entities of the two layers have a linear relationship between the degree and betweenness similarities. If the distribution of points is skewed toward the x-axis, the entities from different networks have similar local roles compared with global roles. If the distribution is skewed toward the y-axis, the entities from various networks have similar global roles compared with local roles.

The procedure of tomographic content analysis consists of the following steps: (a) network construction, (b) calculation of degree and betweenness centralities for each node, (c) node clustering into several modules, (d) node chunking by layers, (e) layer decomposition into four quadrants, (f) descriptive analysis of modules by layers, and (g) similarity analysis between layers.

Results

Data Set

For our study, three keywords—"Parkinson's disease," "PD," and "Parkinson"—were the initial entities of a query for constructing the PBN; 48,885 articles were used. The proteins and genes in the current pathways of Parkinson's disease from PDMAP (Fujita et al., 2014) were selected as the initial entities for the SBN; 87,330 papers were used. The initial entities for the TBN are the same as those of the primary knowledge structure, but expanded with a citation relationship; 110,117 papers were used. All the documents are from MEDLINE, including the corresponding initial entities. In the case of the tertiary structure, additional papers cited by the initial papers found by the tertiary knowledge entities were added as a tertiary set of papers. The cited papers were collected from PMC in MEDLINE. We developed a Java program to collect publications in XML, and to process and extract entities (i.e., MeSH terms given by MEDLINE) from them. In this study we pruned edges with weights lower than 10 to reduce complexity. Consider that our approach is to slice layers from top to bottom to cover all the terms within the left and focused edges, whereas other studies on network comparison by node level have covered only top-ranked nodes. The left ones can reveal the

DOI: 10.1002/asi

difference between networks and are enough for comparative evaluation. We applied network algorithms in Gephi (https://gephi.org/) to calculate centralities and modules.

Statistical Properties

Statistical properties focus on the distribution of degree and betweenness centrality without matching terms. We use these properties to compare two networks by measuring their similarity. Because the variances of each network centrality are different, we apply the two-way *t*-test (Welch's *t*-test) to examine whether pairs of the three bibliometric networks are statistically similar.

In degree, not only primary and secondary (*p* value: .03) but also primary and tertiary (*p* value: .03) are significantly different. In betweenness centrality, no pair could not reject the null hypotheses. Analysis of variance (ANOVA) among three networks results in an insignificant difference (*p* value: .07). The results of tests, applied to the networks without setting a threshold, are similar but ANOVA results in a significant discrepancy (*p* value: .0005). All of the significance levels of the tests are fixed to 95%. Thus, we conclude that at least one network is different from the others in degree but they are similar in betweenness centrality.

Network Properties

We compare the three networks' properties: the number of nodes, the number of edges, average degree, diameter, radius, average path length, the number of shortest paths, density, modularity, and the number of communities. The properties are utilized when investigating a large social and physical network (Barrat, Barthelemy, Pastor-Satorras, & Vespignani, 2004).

The number of nodes for the PBN, SBN, and TBN is 4,065, 7,296, and 6,857, respectively. In terms of entity size, the SBN is the largest, the TBN is the second largest, and the PBN is the smallest. Entity size implies the diversity of nodes for a bibliometric network. From the perspective of terminology, the SBN contains a wide variety of concepts with regard to PD, whereas the PBN involves a relatively narrow scope of terms. The number of edges for the PBN, SBN, and TBN is 102,341, 186,063, and 216,195, respectively. With regard to the increase in entity size, the SBN is the largest, but its edges are second in size. This indicates that the SBN is the sparsest, whereas the PBN is the densest among the three networks. The other network properties confirm this tendency.

The average degrees of the three bibliometric networks indicate that, on average, a node on the PBN has 50.352 neighbors, a node on the SBN has 51.004 neighbors, and a node on the TBN has 63.058 neighbors. It seems PBN and SBN are similar. However, in terms of the number of nodes and edges, the size of SBN is approximately twice as large as the size of PBN. Besides, previously, statistical properties also showed the degree distributions of the two are dissimilar.

The SBN is the largest but sparsest network. Its diameter is 5 as the PBN and longer than the TBN's diameter, 4. But its radius and average path length are the longest. The radius for the PBN, SBN, and TBN is 1, 3, and 2, respectively. The average path lengths of the three networks are 2.06034, 2.11314, and 2.06113, respectively. Path length refers to the number of paths between two nodes. The SBN has the lowest density (0.007), followed by the TBN (0.009) and the PBN (0.012).

Modularity and average clustering coefficient show the same characteristics, but they consider the subgroups, e.g., communities, in each network. The number of communities for the PBN, SBN, and TBN is 8, 6, and 5, respectively. The SBN modularity is the lowest (0.211), followed by the TBN (0.228) and the PBN (0.236). Modularity is the density of edges within communities in networks. This confirms that the SBN is the largest but sparsest network, and the PBN is the smallest but densest.

Tomographic Properties

This section examines whether, and how, the content of the three bibliometric networks is different from one another. Tomographic analysis allows us to examine a network topology at different levels. Appendix A (a–c) visualizes the three bibliometric networks built by the traditional approach, which includes the top 100 nodes and represents the degree centralities of the nodes in node sizes. These figures show nodes in only one metric, in this case degree, or the local importance within the top 100 nodes for each network. To identify similarities and differences between networks systematically, more nodes and metrics should be considered. To this end, we adopt the combinatory approach of the degree-betweenness plane shown in the Appendices. This plane can show the overall pattern of each network where the cluttering of nodes makes it difficult to compare two networks. Analyzing the degree-betweenness plane per layer helps identify the fine-grained properties of the three networks. The appendices list 5 to 10 layers for the three networks (http://informatics.yonsei.ac.kr/tsmm/kh_lee/appendix.pdf).

Tomographic Content Analysis

We conducted tomographic content analysis for the three bibliometric networks to evaluate the local and global position of an entity at a layer. Appendix A shows the overall position of MeSH entities in the PBN. Figure 3a–d shows the position of MeSH entities at layers from 1 to 4. We generate 10 layers for each network and show them in the Appendices.

Primary Bibliometric Network (PBN)

In this section, we describe the PBN from a tomographic perspective on the degree-betweenness plane at each layer. The PBN has "humans" as the most popular entity locally and globally, as shown in Figure 3a. The entity frequently appears and is applied as a mediator between other concepts. Both the SBN and TBN share this entity as the most popular.

DOI: 10.1002/asi

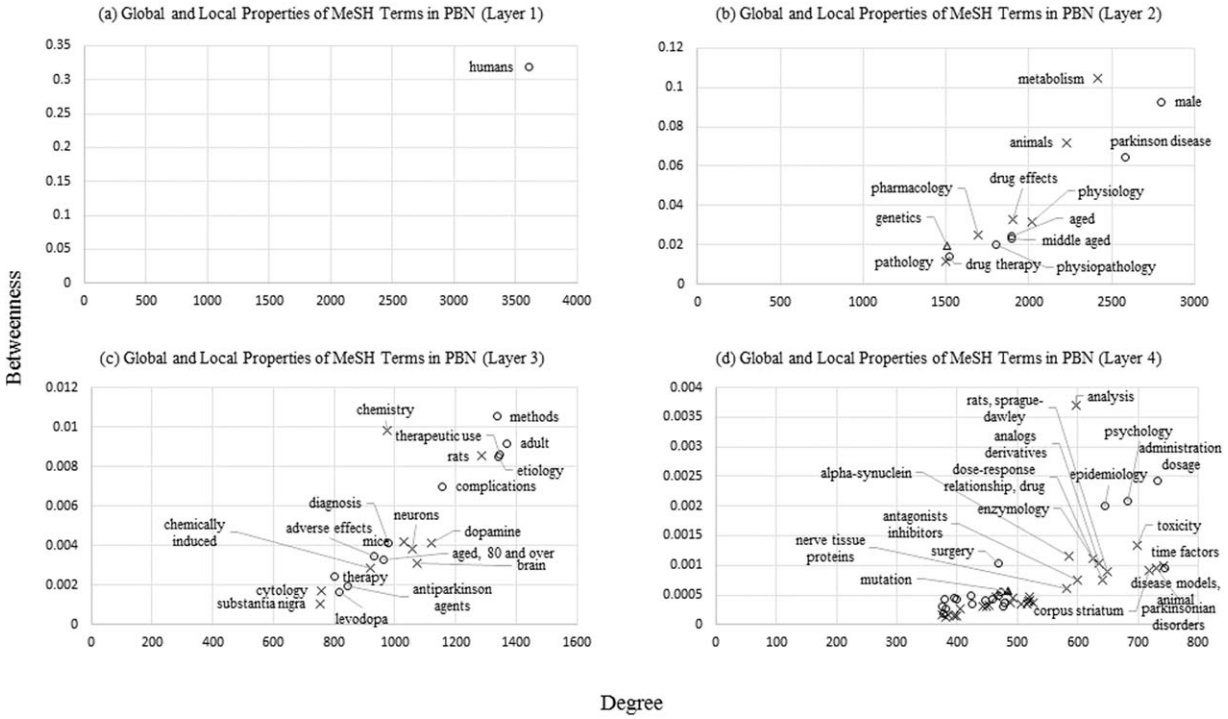

FIG. 3. Global and local properties for MeSH terms in the top four degree-betweenness layers for PBN of PD: (a) layer 1, (b) layer 2, (c) layer 3, (d) layer 4.

The PBN consists of eight modules detected by the community detection technique, and these eight modules are distributed over the 10 layers. Of these eight modules, three are major and the remainder are minor. One major module in the second layer, which includes "humans" and is symbolized with a circle, contains "male," "Parkinson's disease," "aged," "middle aged," "physiopathology," and "drug therapy." As we confirm in Figure 3a–d, this module involves the descriptions, symptoms, and therapy for PD. Another major module in the second layer, which contains "metabolism" and "animals" and is symbolized as crosses, involves "physiology" and "pharmacology." The other major module—which includes "genetics," occurs in the second layer, and is symbolized as triangles in Figure 3—represents genes, proteins, and relevant notions.

Tomographic content analysis indicates that the PBN describes PD with an emphasis on genetics, animal experiments for cure, and physiopathological symptoms. The module that involves "methods" in the third layer is linked to the module related to treatments, such as administration dosage, time factors, case–control studies, neuropsychological tests, and treatment outcome. The module that includes "brain" in the third layer is associated with a module with particular genes (e.g., "antagonist inhibitors," "alpha-synuclein, nerve tissue proteins," "immunohistochemistry," "tyrosine 3-monooxygenase," and "neuroprotective agents"). This module leads us to infer that there are various approaches to investigating PD. It may indicate that PD is on nerve cells, and their generation and degeneration are the main interest. The PBN focuses on disease. It is supported by three main modules conceptualized as "genetics," "animal experiments," and "human symptoms." Modules other than "genetics" are large. The minor modules are on the seventh layer. One prominent module is on history. In the 10th layer, three minor modules are detected. They contain "chromogranins," "allyl compounds" and "butylamines," and "rain."

The terms in HH that are regarded as important entities locally and globally for each layer are associated with human PD patients, and those terms include "male," "Parkinson's disease," "therapeutic use," "etiology," "complications," "administration dosage," "psychology," "statistics numerical data," "tomography, x-ray computed," and "electrocardiography." They are also linked with pharmacology and include "metabolism," "animals," "rats," "chemistry," "analysis," "immunology," "iron," "anti-inflammatory agents, non-steroidal," and "receptor, serotonin, 5-ht1a." In the lower layers, other topical terms, such as "history, 19th century," "ethics," "rain," and "chromogranin," are detected as important. However, genetic terms are excluded from HH.

Overall, the PBN focuses on the PD field from the perspective of pharmacology. Bridging genetic terms are discovered, but their positions are insufficiently high. The relevant history and ethics are also treated as important in the lower layers. From a knowledge discovery perspective, given that the PBN covers principle entities in PD, it serves as the basic network for exploring the PD field.

Secondary Bibliometric Network (SBN)

In this section we describe the SBN from a tomographic perspective with degree-betweenness layers. The SBN has

DOI: 10.1002/asi

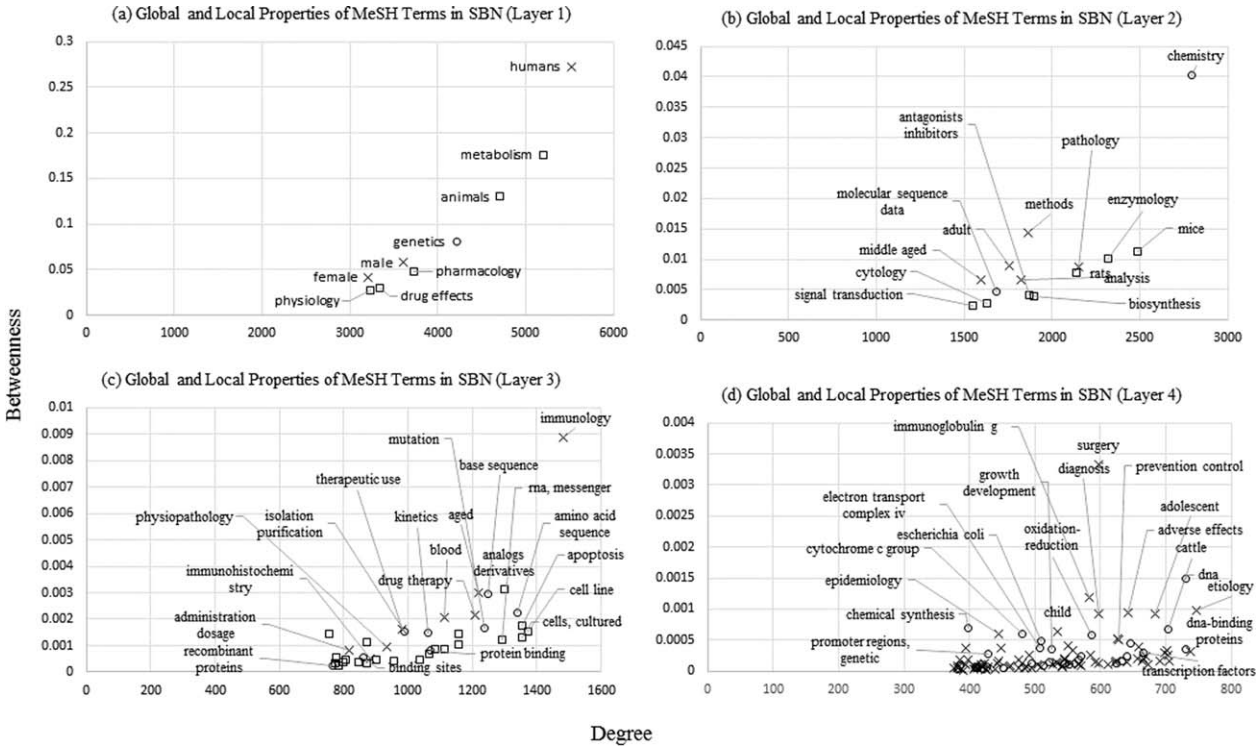

FIG. 4. Global and local properties for MeSH terms in the top four degree-betweenness layers for SBN of PD: (a) layer 1, (b) layer 2, (c) layer 3, (d) layer 4.

three major modules; "humans" is the most salient entity locally and globally in the first layer.

The SBN is divided into six modules throughout the 10 layers; three are major and the remainder are minor. One major module represents cell biology relevant to pharmacology, and it is plotted with rectangles. The module contains "metabolism," "animals," "pharmacology," "drug effects," and "physiology" in the first layer, and "mice," "enzymology," "rats," "biosynthesis," "antagonist's inhibitors," "cytology," and "signal transduction" in the second layer. As we can confirm in Figure 4a–d, this module involves general pharmacological terms. Another major module, which includes "humans" and is symbolized with crosses, contains "male" and "female" in the first layer and "methods," "pathology," "analysis," and "adult" in the second layer. The other major module, which includes "genetics," occurs in the first layer, and is symbolized with circles in Figure 4. The module contains "chemistry" and "molecular sequence data" in the second layer. We can confirm that the SBN contains broader concepts than the PBN, which is centered on PD.

The minor modules are shown in the sixth layer. There are several interesting findings observed in the SBN. First, the names of nations appear in two different modules. One module has "humans" in the first layer, and it includes "United States," "China," and "Europe" in the sixth, seventh, and eighth layers, respectively. The module contains demographic, social, and organizational terms, such as "economics," in the seventh and eighth layers. Second, a module of dentistry terms is established in the sixth layer (e.g., "surface properties," "materials testing," and "denture bases"). Third, a module for healthcare in developing countries is formed in the ninth layer and the 10th layer (e.g., "developing countries," "health planning," "Asia," "family planning services," "organization and administration," and "delivery of health care").

Overall, the SBN focuses on the peripheral entities associated with PD. Compared with the PBN, a wide range of biological concepts and macroscopic notions relevant to public health are increased, whereas the descriptive entities of PD are reduced. For example, the terms for "dentures" is increased. PD patients are often old and have trembling symptoms, so that prosthodontic conditions should be considered. Such consideration may be indirectly relevant to PD, but it can lead us to contemplate problems derived from PD. From the perspective of knowledge discovery, the SBN can illuminate new linkages among PD entities.

Tertiary Bibliometric Network (TBN)

We describe the TBN from a tomographic perspective with degree-betweenness layers. The TBN has two modules where entities such as "humans," "male," and "female" form one module symbolized with a circle, and entities such as "metabolism," "animals," "physiology," "genetics," "pharmacology," and "drug effects" form another module symbolized with a triangle, as shown in Figure 5a–d. The entity "humans" is the most salient locally and globally in the first layer.

In the second layer, there are two modules that are the successors of the two modules in the first layer. One consists

DOI: 10.1002/asi

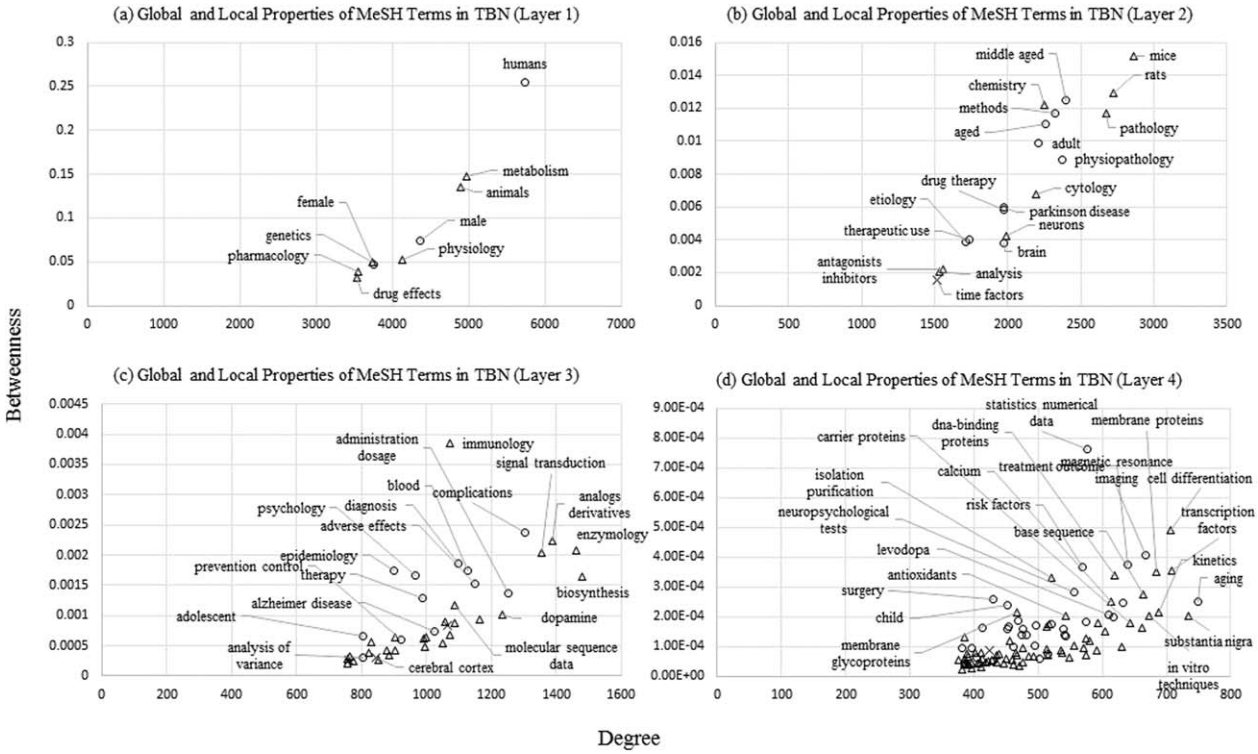

FIG. 5. Global and local properties for MeSH terms in the top four degree-betweenness layers for TBN of PD: (a) layer 1, (b) layer 2, (c) layer 3, (d) layer 4.

of "middle aged," "physiophathology," "methods," "aged," "adult," "Parkinson's disease," "brain," "drug therapy," "therapeutic use," and "etiology." The other one has "mice," "rats," "pathology," "chemistry," "cytology," "neurons," "antagonistic inhibitors," and "analysis." A new module that consists of "time factors" is generated in this layer. In the third layer, the module that contains "humans" in the first layer is expanded and forms a module that includes "complications," "administration dosage," "adverse effects," "blood," "diagnosis," and "Alzheimer's disease." Overall, these results show that the TBN covers more general entities associated with PD than the SBN, but those TBN entities are not as core as those of the PBN.

The TBN is divided into five modules. The two major modules are on PD and the experiments for investigating PD cause and effect. The terms "humans," "male," and "female" appear on the first layer. The second layer contains module terms such as "middle aged," "physiopathology," "methods," "Parkinson's disease," "drug therapy," and "etiology."

The second one is on experiments on etiology, including physiological, genetic, and pharmacological perspectives. The remaining three minor modules are on "brain," "biochemistry," and "kidney cancer." The first one is on brain. The first entity of the module that appears on the high level is "time factors" on the third layer. The subsequent entities are "chemically induced," "cerebral cortex," and "motor activity" on the fourth layer; "electric stimulation" on the fifth layer; "radiation effects," "action potentials," "serotonins," "cerebellum," "dopamine agents," "maze learning," "cocaine," and relevant terms on the sixth layer. The second one is on biochemistry. "Biochemistry" is located on the 10th layer. The third one is on kidney cancer. "Kidney neoplasms" is positioned on the 10th layer. One cohort study reported that males with Parkinson's disease had 16% higher risk of kidney cancer than those without (Ong, Goldacre, Goldacre, 2014).

The entities of the major modules in the HH sections are "humans," "metabolism," and "animals" on the first layer; "mice," "rats," "pathology," "middle aged," "chemistry," "methods," "adult," "physiopathology," and "cytology" on the second layer; "immunology," "complications," "signal transduction," "analogs derivatives," "enzymology," and "biosynthesis" on the third layer; and "cell differentiation transcription factors," membrane proteins," "magnetic resonance imaging," and "statistics numerical data" on the fourth layer. We confirm that PD and the relevant experimental knowledge entities are detected as the major modules, but the relevant knowledge entities are treated as important.

When comparing the first layer of the three networks, the PBN contains the primary entities associated with Parkinson's disease, whereas the SBN contains more general bioentities, such as genetics, humans, and animal drug experiments. The TBN lies between PBN and SBN in terms of the topical matter of entities. In the right upper corner, there are popular entities such as "female," "drug therapy," "age," and "conditions during a disease state," and they are treated as important in the PBN. The SBN and TBN share similar entities, such as "genetics," "metabolism," "animals," "pharmacology," "drug effects," and "physiology." However,

DOI: 10.1002/asi

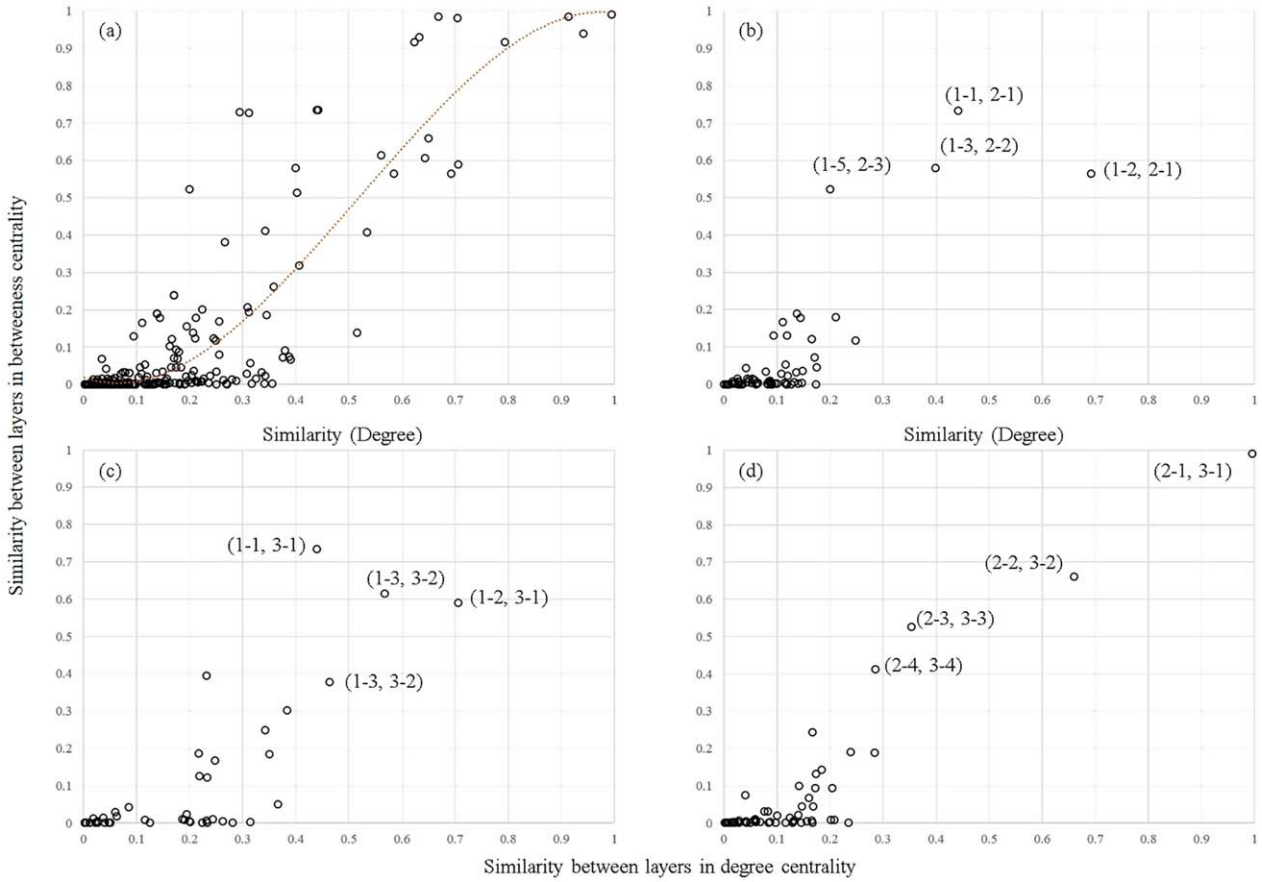

FIG. 6. Similarity plots between layers from betweenness and degree centralities perspective: (a) pairwise comparisons between three networks, (b) pairwise comparisons between PBN and SBN, (c) pairwise comparisons between PBN and TBN, (d) pairwise comparisons between SBN and TBN.

the SBN separates "genetics" from other entities, whereas the TBN has "genetics" along with "metabolism," "animals," "pharmacology," "drug effects," and "physiology."

Considering the first layer only, the entities among the three networks are similar. The theme of the first layer of the PBN is Parkinson's disease. The first layer of the SBN can be abstracted as genetics, subjects, and pharmacological animal experiment. That of the TBN can be summarized as subjects, and pharmacological and genetic animal experiments. Focus on the current knowledge on PD might benefit from the PBN. Focus on the new knowledge and links with other fields associated with PD might benefit from the SBN. Focus on the current knowledge on PD and relevant studies might benefit from the TBN.

Discussion

We performed descriptive analyses on the PBN, SBN, and TBN by layer. Further, we compared the networks by layer and modules. This was accomplished by plotting the content similarities of degree and betweenness centralities between the layers of the networks on a two-dimensional space.

Figure 6a shows the content similarity among the three networks. Overall, the content similarity between layers is low, in that most similarity values are plotted in the bottom left corner. Figure 6b illustrates the similarities between the PBN and SBN layers. (1-5, 2-3) denotes the position that corresponds to the comparison results between the fifth PBN and the third SBN layers. The two layers are more similar in terms of global than local impact. "Immunology" is located in the HH sections of the two layers. Genetic and cell biological terms frequently appear in the SBN. Resemblance in degree centrality is relatively low. This implies that the impactful entity of the two layers is similar, but the component terms differ. Another example is (1-2, 2-1). The two layers are highly similar and have three similar modules that describe animal experiments that involve pharmacology, genetics, and human patients. We see that similar topics are located on different levels in the two networks. Figure 6c depicts the similarities between the PBN and TBN layers. Similar to Figure 6b, (1-1, 3-1) implies that the main entities coincide with "humans." (1-2, 3-1) and (1-3, 3-2) indicate the topological differences between the PBN and TBN. Similar content in the local and global effects of the networks is located in different levels. Figure 6d shows the equivalence between the SBN and TBN layers. Interestingly, the two networks resemble each other from the first to fourth levels, but the extent of equivalency decreases.

Table 1 lists the pairwise similarities between two of the layers from the three networks for 10 layers. The cell on the

DOI: 10.1002/asi

TABLE 1. Pairwise cosine similarity between two of the PBN, SBN, and TBN in degree and betweenness centralities.

	PBN vs. SBN		SBN vs. TBN		PBN vs. TBN	
Layer	Degree similarity	Betweenness similarity	Degree similarity	Betweenness similarity	Degree similarity	Betweenness similarity
1	0.734	0.442	0.99	0.995	0.734	0.439
2	0.028	0.109	0.661	0.66	0.185	0.35
3	0.043	0.042	0.526	0.354	0.186	0.217
4	0.178	0.144	0.189	0.284	0.126	0.218
5	0.166	0.111	0.412	0.286	0.168	0.248
6	0.189	0.138	0.008	0.202	0.023	0.194
7	0.031	0.137	0.008	0.209	0.004	0.199
8	0.003	0.115	0.068	0.16	0.002	0.201
9	0	0.083	0.001	0.167	0.011	0.187
10	0	0.174	0.001	0.235	0.006	0.231

first row and column is the content similarity between the layers in the same level (layer 1) of the PBN and SBN by degree centrality. Layer 1 for both the PBN and SBN has content similarity of 0.734 by degree centrality. As indicated in Table 1, the results confirm that the similarity between layers decreases with deeper layers. From layers 1 to 5, the layers become more different. The first layers for the SBN and TBN are verified as the most similar in both degree and betweenness centralities. Table 1 also indicates support for a hypothesis where the networks are similar with one another in higher layers, but become dissimilar in lower layers. Subsequently, tomographic network analysis allows us to dissect networks in levels to identify the networks thoroughly. The conventional approach to concentrate on the

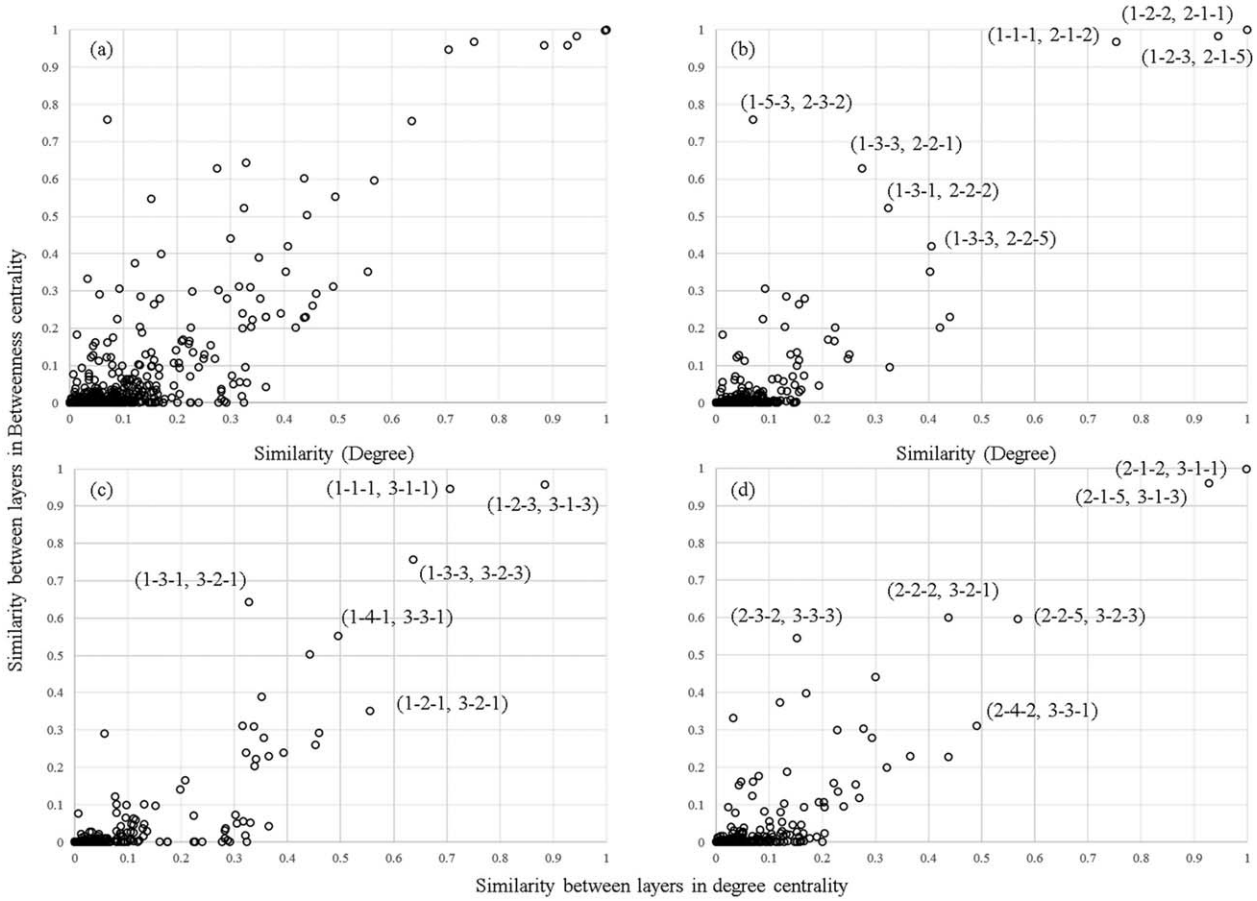

FIG. 7. Similarity plots between layer modules from betweenness and degree centralities perspectives: (a) pairwise comparisons between three networks, (b) pairwise comparisons between PBN and SBN, (c) pairwise comparisons between PBN and TBN, (d) pairwise comparisons between SBN and TBN.

DOI: 10.1002/asi

central nodes and edges might lose the distinctive features of a network, which is confirmed by the results of analyzing statistical properties. This makes the network general, so that network comparison becomes less useful. The results show that the SBN and TBN are significantly similar only when comparing the distributions of centralities in the first layers of the two networks. Counterintuitively, the entities in the literature that contain a lower ontological level to PD have similar degree and betweenness centralities as the entities from citation-related literatures. The SBN is built by the entities extracted from articles that contain genes/proteins ontologically related to PD. The TBN is established with entities extracted from PD papers and their cited papers.

Figure 7 shows the possible deeper understanding of the resemblance and unlikeness of the three networks. Figure 7a shows the total comparisons between all network modules by layers. Figure 7b–d is the corresponding plot for Figure 6b–d. Comparing Figures 6 and 7, we have another metric for identification of the module of a network layer. (1-2-2, 2-1-1) defines the likeness between module 2 of the PBN second layer and module 1 of the SBN first layer. The two modules contain "metabolism," "animals," "pharmacology," "drug effects," and "physiology," and their centralities in degree and betweenness are almost the same. From Figure 7, we can closely associate or affiliate two networks.

The present study finds that bibliometric networks on a certain topic have distinctions. The PBN is a conventional network that scholars often use, and has the advantage of clear understanding of a certain topic, for example, PD. The PBN focuses on the general description of PD and the relevant pharmacology. However, it might not be adequate for discovering innovative links between knowledge entities for new knowledge discovery. The SBN benefits from a wide range of weak ties to a certain topic. The SBN connects PD with the side effect of a denture problem and with public health. It extends the knowledge structure of PD, similar to the PBN, to cover more innovative associations compared with the PBN. The knowledge structure of PD is part of the SBN, whereas it is the center of the PBN. The SBN is limited in its ability to provide profound knowledge comprehension on the target topic. The TBN has features that are similar to the PBN, but it is linked with closer concepts to PD than the SBN.

Conclusion

As science becomes more interdisciplinary and its repositories are enlarged, bibliometric network analysis gains more attention from the scholarly community. Such analysis allows us to summarize a certain topic and generate new hypotheses and findings without requiring humans to process the entire accumulated knowledge.

Understanding the features of various bibliometric networks on a certain topic is important in summarizing a general outline of the topic and in discovering new linkages between the current knowledge entities, which allows scholars to choose an appropriate bibliometric network construction scheme to achieve their goals. The significance of the understanding of network properties grows larger as knowledge accumulates to a point where it overwhelms human cognition.

The present study evaluated three bibliometric networks in the field of PD. The networks were built as knowledge structures and categorized as the PBN, SBN, and TBN according to the initial querying entities. Each network was analyzed by layers of the degree-betweenness plane that represents the local and global impacts of knowledge entities on the corresponding network.

Traditionally, network comparison performed by evaluating statistical properties, such as the distributions of degree centrality, has been well received. Network analysis over a small number of top nodes has been sufficient for network comparison. The findings of our study, however, indicate that such traditional tools are insufficient for evaluating bibliometric networks in detail and distinguishing one from another.

The contributions of the present study are three-fold. First, we categorized various bibliometric networks into three types according to the initial query selection. Second, we proposed a comparative evaluation approach, called tomographic content analysis, for analyzing the three networks in detail. Third, the findings from comparing the PBN, SBN, and TBN suggest that the PBN is adequate for understanding the current knowledge of a certain topic, whereas the SBN helps discover new knowledge and linking to other fields associated with the topic, and the TBN is best suited for overviewing the current knowledge of a topic and its relevant studies.

This study restricted the knowledge entities to MeSH terms. Although MeSH terms are sufficient for inducing a general outline of a certain topic, we relaxed the controlled terms and included text in titles and abstracts. We limited the axes of a layer to degree and betweenness centralities. These are the most popular indicators for representing the local and global impacts of a node on other nodes. However, investigating other measurements could be worthwhile.

As a follow-up study, we plan to incorporate authorships and affiliations into building the bibliographic networks. This could provide hints for new scientific and synergic collaboration. For new knowledge discovery, we might add various types of knowledge entities detected by named entity recognition to expand the scope of bibliometric networks.

Acknowledgments

This work was supported by the Bio-Synergy Research Project (NRF-2013M3A9C4078138) of the Ministry of Science, ICT and Future Planning through the National Research Foundation, as well as the Yonsei University Future-leading Research Initiative of 2015 (2015-22-0119).

DOI: 10.1002/asi

References

Al-Kindi, S., Al-Juhaishi, T., Haddad, F., Taheri, S., & Khalil, C.A. (2015) Cardiovascular disease research activity in the Middle East: A bibliometric analysis. Therapeutic Advances in Cardiovascular Disease, 9, 70–76.

Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. (2004) The architecture of complex weighted networks. Proceedings of the National Academy of Sciences of the United States of America, 101, 3747–3752.

Blockmans, W., Engwall, L., & Weaire, D. (Eds.). (2014) Proceedings from the Bibliometrics: Use and Abuse in the Review of Research Performance Symposium in Stockholm, Sweden. Portland, ME: Portland Press.

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, 2008(10), P10008.

Bramness, J.G., Henriksen, B., Person, O., & Mann, K. (2013) A bibliometric analysis of European versus USA research in the field of addiction. Research on alcohol, narcotics, prescription drug abuse, tobacco and steroids 2001-2011. European Addiction Research, 20, 16–22.

Cantos-Mateos, G., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., & Zulueta, M.A. (2012) Stem cell research: Bibliometric analysis of main research areas through KeyWords plus. Aslib Proceedings. 64(6), 561–590.

Chen, H., Wan, Y., Jiang, S., & Cheng, Y. (2014) Alzheimer's disease research in the future: Bibliometric analysis of cholinesterase inhibitors from 1993 to 2012. Scientometrics, 98, 1865–1877.

Cheng, T., & Zhang, G. (2013) Worldwide research productivity in the field of rheumatology from 1996 to 2010: A bibliometric analysis. Rheumatology, 52, 1630–1634.

Ding, F., Jia, Z., & Liu, M. (2016) National representation in the spine literature: A bibliometric analysis of highly cited spine journals. European Spine Journal, 25(3), 850–855.

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., & Chambers, T. (2013) Entitymetrics: Measuring the impact of entities. PLoS One, 8, e71416.

Fujita, K.A., Ostaszewski, M., Matsuoka, Y., Ghosh, S., Glaab, E., Trefois, C., ... Balling, R. (2014) Integrating pathways of Parkinson's disease in a molecular interaction map. Molecular Neurobiology, 49, 88–102.

Gu, W., Yuan, Y., Yang, H., Qi, G., Jin, X., & Yan, J. (2015) A bibliometric analysis of the 100 most influential papers on COPD. International Journal of Chronic Obstructive Pulmonary Disease, 10, 667.

Gupta, B.M., & Bala, A. (2013) Parkinson's disease in India: An analysis of publications output during 2002-2011. International Journal of Nutrition, Pharmacology, Neurological Diseases, 3, 254.

Ho, Y.S., Chiu, C.H., Tseng, T.M., & Chiu, W.T. (2003) Assessing stem cell research productivity. Scientometrics, 57, 369–376.

Ho, Y.S., Nakazawa, K., Sato, S., Tamura, T., Kurishima, K., & Satoh, H. (2013) Cisplatin for small cell lung cancer: Associated publications in Science Citation Index Expanded. Oncology Letters, 5, 684–688.

Hoffmann, R., & Valencia, A. (2004) A gene network for navigating the literature. Nature Genetics, 36, 664–664.

Huang, Z. (2013) Mining disease associated biomarker networks from PubMed. In 2013 Seventh International Conference on Systems Biology (ISB) (pp. 15–18). Yellow Mountain, China: IEEE.

Huffman, M.D., Baldridge, A., Bloomfield, G.S., Colantonio, L.D., Prabhakaran, P., Ajay, V.S., ... Prabhakaran, D. (2013) Global cardiovascular research output, citations, and collaborations: A time-trend, bibliometric analysis (1999–2008). PLoS One, 8, e83440.

Jenssen, T.K., Lægreid, A., Komorowski, J., & Hovig, E. (2001) A literature network of human genes for high-throughput analysis of gene expression. Nature Genetics, 28, 21–28.

Kostoff, R. N. (2014) Literature-related discovery: common factors for Parkinson's Disease and Crohn's Disease. Scientometrics, 100, 623–657.

Krishnamoorthy, G., Ramakrishnan, J., & Devi, S. (2009) Bibliometric analysis of literature on diabetes (1995-2004). Annals of Library and Information Studies, 56, 150.

Lee, D., Kim, W.C., Charidimou, A., & Song, M. (2015) A bird's-eye view of Alzheimer's disease research: Reflecting different perspectives of indexers, authors, or citers in mapping the field. Journal of Alzheimer's Disease: JAD, 45, 1207–1222.

Lewin, H.S. (2008) Diabetes mellitus publication patterns, 1984–2005. Journal of the Medical Library Association: JMLA, 96, 155.

Lewison, G., Rippon, I., De Francisco, A., & Lipworth, S. (2004) Outputs and expenditures on health research in eight disease areas using a bibliometric approach, 1996–2001. Research Evaluation, 13, 181–188.

Li, T., Ho, Y.S., & Li, C.Y. (2008) Bibliometric analysis on global Parkinson's disease research trends during 1991–2006. Neuroscience Letters, 441, 248–252.

Li, Z., Qiu, L.X., Wu, F.X., Yang, L.Q., Sun, Y.M., Lu, Z.J., & Yu, W.F. (2012) Assessing the national productivity in subspecialty critical care medicine journals: A bibliometric analysis. Journal of Critical Care, 27, 747-e1.

Natarajan, J., Berrar, D., Dubitzky, W., Hack, C., Zhang, Y., DeSesa, C., ... Bremer, E.G. (2006) Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. BMC Bioinformatics, 7(1), 373.

Ong, E.L., Goldacre, R., & Goldacre, M. (2014) Differential risks of cancer types in people with Parkinson's disease: A national record-linkage study (link is external). European Journal of Cancer, 50, 2456–2462.

Ponce, F.A., & Lozano, A.M. (2011) The most cited works in Parkinson's disease. Movement Disorders, 26, 380–390.

Rodrigues, P.S., Fonseca, L., & Chaimovich, H. (2000) Mapping cancer, cardiovascular and malaria research in Brazil. Brazilian Journal of Medical and Biological Research, 33, 853–867.

Royle, P., & Waugh, N. (2015) Macular disease research in the United Kingdom 2011–2014: A bibliometric analysis of outputs, performance and coverage. BMC Research Notes, 8, 1–8.

Schäfer, A., Hiemke, C., & Baumann, P. (2015) Consensus guideline for therapeutic drug monitoring in psychiatry (2004): Bibliometric analysis of citations for the period 2004–2011. Nordic Journal of Psychiatry, 1–6.

Song, M., Han, N.G., Kim, Y.H., Ding, Y., & Chambers, T. (2013) Discovering implicit entity relation with the gene-citation-gene network. PLoS One, 8, e84639.

Song, M., Heo, G.E., & Lee, D. (2015) Identifying the landscape of Alzheimer's disease research with network and content analysis. Scientometrics, 102, 905–927.

Sorensen, A.A., Seary, A., & Riopelle, K. (2010) Alzheimer's disease research: A COIN study using co-authorship network analytics. Procedia-Social and Behavioral Sciences, 2, 6582–6586.

Sorensen, A.A., & Weedon, D. (2011) Productivity and Impact of the Top 100 Cited Parkinson's Disease Investigators since 1985. Journal of Parkinson's Disease, 1, 3–13.

Stapley, B.J., & Benoit, G. (2000) Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in Medline abstracts. Pacific Symposium on Biocomputing, 5, 529–540.

Swanson, D.R. (1986) Undiscovered public knowledge. The Library Quarterly, 103–118.

Sweileh, W.M., Sa'ed, H.Z., Al-Jabi, S.W., & Sawalha, A.F. (2014) Bibliometric analysis of diabetes mellitus research output from Middle Eastern Arab countries during the period (1996–2012). Scientometrics, 101, 819–832.

Tesio, L., Gamba, C., Capelli, A., & Franchignoni, F.P. (1995) Rehabilitation: The Cinderella of neurological research? A bibliometric study. The Italian Journal of Neurological Sciences, 16(6), 473–477.

Ugolini, D., Neri, M., Cesario, A., Marazzi, G., Milazzo, D., Volterrani, M., ... Pasqualetti, P. (2013) Bibliometric analysis of literature in cerebrovascular and cardiovascular diseases rehabilitation: Growing

numbers, reducing impact factor. Archives of Physical Medicine and Rehabilitation, 94, 324–331.

Ugolini, D., Puntoni, R., Perera, F.P., Schulte, P.A., & Bonassi, S. (2007) A bibliometric analysis of scientific production in cancer molecular epidemiology. Carcinogenesis, 28, 1774–1779.

Uthman, O.A., Wiysonge, C.S., Ota, M.O., Nicol, M., Hussey, G.D., Ndumbe, P.M., & Mayosi, B.M. (2015) Increasing the value of health research in the WHO African Region beyond 2015—Reflecting on the past, celebrating the present and building the future: A bibliometric analysis. BMJ Open, 5, e006340.

Van Leeuwen, T.N., Van der Wurff, L.J., & Van Raan, A.F.J. (2001) The use of combined bibliometric methods in research funding policy. Research Evaluation, 10, 195–201.

Blockmans, W., Engwall, L., & Weaire, D. (Eds.). (2014) Bibliometrics: Use and Abuse in the Review of Research Performance: Proceedings from a Symposium Held in Stockholm, 23–25 May 2013. Portland Press.

Yang, J., Vannier, M.W., Wang, F., Deng, Y., Ou, F., Bennett, J., ... Wang, G. (2013) A bibliometric analysis of academic publication and NIH funding. Journal of Informetrics, 7, 318–324.

Yu, Q., Ding, Y., Song, M., Song, S., Liu, J., & Zhang, B. (2015) Tracing database usage: Detecting main paths in database link networks. Journal of Informetrics, 9, 1–15.

Zyoud, S.H., Al-Jabi, S.W., Sweileh, W.M., & Awang, R. (2014) A bibliometric analysis of research productivity of Malaysian publications in leading toxicology journals during a 10-year period (2003–2012). Human & Experimental Toxicology, 33(12), 1284–1293.

DOI: 10.1002/asi