Knowledge-Based Systems 108 (2016) 168–178

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Cannibalism in medical topic networks

Suhyun Chae, Aviv Segev*, Uichin Lee

Department of Knowledge Service Engineering, KAIST, Daejeon, 305-701, South Korea

ARTICLE INFO

Article history:
Received 30 October 2015
Revised 8 May 2016
Accepted 9 May 2016
Available online 11 May 2016

Keywords:
Research publication activity
Network evolution
Keyword extraction
Knowledge structure
Medical domain

ABSTRACT

Analyzing research activities over time can give insight into the research trend and knowledge structure of a domain. Research publication activity of a topic can be measured by a network of keyword terms and their relations in the specific area. The paper analyzes medical topic networks to interpret how clusters and keyword terms change over time. Keywords are extracted from 9730,671 research publications of twenty medical topics over 40 years. Experiments show there is cannibalism which occurs when one cluster is consumed into other clusters of medical topic networks in 50% of the medical topics analyzed. The decrease of modularity values of cannibalism topics shows that research topics collaborate actively and that multidisciplinary fields have emerged over time.

1. Introduction

A social network is a network that consists of members in a group as nodes and their relationship as links [1]. Nodes and links can be defined differently in many kinds of networks [2]. A common type of network is a collaboration network. Scientific collaboration networks represent how scientists are connected to other researchers. Scientific publications, usually called co-authorship networks, are one of the most tangible, well-documented forms of scientific collaboration [3].

Generally, the expertise of scientists is examined by their publication activities. Many researches have revealed the structure of scientist communities by analyzing the co-authorship network among scientists. However, not only researchers have collaboration with their publication activities. Like scientists, research topics also have relations.

Like with scientists, the trend of a topic can be examined by its publication activity. There are many scientific papers in a topic and many keywords in a paper. Terms are defined as the core keywords in an article. The collection of terms represents the behavior of a topic. Therefore, research publication activity is defined as the number of publications on the topic analyzed by term occurrences.

The main goal of this research is to show a common pattern of changing scientific research networks among topics. Research publication data, including publication year, journal assigned keywords, and author assigned keywords, were collected from Web of Science and PubMed. Medical topics are selected in this paper because they have a more defined structure compared to other domains. Keywords are extracted in each topic from articles to make term networks. Keywords that are less related to the topic are eliminated. Temporal networks were created with the remaining keywords for each topic and each time period.

This paper suggests analyzing term networks in a medical topic to learn how the topic is changing over time. In the term networks, nodes are topic-related keywords and two nodes are linked if they both appeared in a single article at least once. The medical topic networks were analyzed to reveal how they are changing over time and to find a common characteristic in the networks. 1,017,486 keywords extracted from 9,730,671 research publications of twenty medical topics were analyzed. The main finding from the research is that in many topics nodes are being merged to create a large component as time passes. Network values are compared to understand when this phenomenon appears.

The rest of the paper is organized as follows. The next chapter reviews the related works. Section 3 describes methods used for term network analysis in medical topics. Section 4 presents the experiments and results on term networks. Section 5 discusses the results and limitations of this research. Finally, Section 6 concludes the paper and suggests future work.

2. Literature review

A social network is a collection of people who are connected by their social relationship [4]. There are many types of social networks, and the interest in networks spans multiple fields such as social sciences, physics, epidemiology, and biology [5]. Social Network Analysis (SNA) has been widely used to understand creativity [6], innovation [7], job performance [8], management consulting

* Corresponding author.
E-mail address: aviv@kaist.edu (A. Segev).

http://dx.doi.org/10.1016/j.knosys.2016.05.017
0950-7051/

[9], promotion [10], turnover [11], and unethical behavior [12] in management research. It is also used in life sciences [13], electrical engineering [14], computer science [15], ecosystems [16], biology [17], and various other fields.

Social network analysis has been considered an approach for analyzing patterns of relationships and interactions between social actors to discover underlying social structure [18]. The social actors include directors of companies [19], organizational behavior [20], inter-organizational relations [21], and computer-mediated communications [22].

One of the social networks is co-authorship networks: nodes are paper authors, joined by edges if they have written at least one paper together [23]. Researchers interact not only to communicate research activities but also to collaborate with each other to produce research and co-author research results [24]. Scientific collaboration has seen huge growth in recent decades, with research groups becoming the minimum unit of the scientific system in many areas [25]. Since collaboration has the potential to promote research activity, productivity, and impact [18], there is a positive correlation between collaboration and co-authorship [26].

Newman [4] analyzed seven scientific collaboration networks and found that networks are small worlds. A small world is any pair of people of a network connected to each other through one or more mutual acquaintances [27]. Kim and Perez [3] analyzed collaboration networks in the industrial ecology research domain. The research suggests that co-authorship maps increase scientific productivity and that the multidisciplinary field of research is continuously growing. These two researches were conducted to find the hidden structure of scientific networks.

Other researches focused on the role of each node. Wasserman and Faust [2] discovered hub nodes, leaders, gatekeepers, highly connected groups, and patterns of interactions between groups. Kademani and his colleagues concentrated on the publication activity of the individual researcher, while Chandrasekhar [28] and Hodgkin [29] worked to know characteristics of Nobel Prize winners.

Communities are common structures in complex systems [30]. Communities are defined as groups of nodes with dense intra-connections and sparse inter-community links [31]. Community detection examines the structural information of a graph to divide it into partitions, communities [32]. A community detection algorithm suggested by Newman [33] is a widely used method. The method uses modularity as a quality function. Another method developed by Louvain [34] is also commonly used because of its low computational complexity and high performance. This method considers the consolidation ratio when merging communities as well as modularity.

However, community structure is affected by dynamic effects [35]. Holme and Saramaki [36] generalized a randomized model to dealing with dynamic communities of temporal networks. The basic idea of the method is to randomize or reshuffle original event sequences to remove time-domain structure and correlations. Palla et al. [35] evaluated the community evolution by joint networks of time t and $t+1$.

Quattrociocchi et al. [37] suggested the importance of analyzing a dynamic network, not a static network. The paper analyzed the evolution of citation networks and co-authorship networks. The work shows how the network values of citation and co-authorship graphs are changing over time respectively. Graph density and modularity values decrease in both networks.

Michel et al. [38] analyzed a corpus of digitized texts in books as a linguistic approach. The analysis enables the quantitative investigation of cultural trends. The results provide insights about fields as diverse as lexicography, evolution of grammar, and collective memory. They analyzed frequency of keywords by year or region and language. The work performed text analysis instead of network analysis to help understand the knowledge structure of a domain.

The previous research focused on frequency of words used in texts. The research analyzed the occurrences to find cultural phenomena. However, the research did not analyze a network. Other previous papers analyzed social networks and the structure of networks. In the present research, medical topic networks were analyzed with nodes as keywords. The present paper combines the two approaches, research network analysis and analysis of keyword occurrences.

3. Method

The goal of this research is to analyze the evolution of research networks over time. Medical topic networks represent core terms of the topics. The networks consist of keywords as nodes and their relations as links. Keywords of a topic are extracted from a whole set of research publications in the topic. Extracted keywords are examined by their relevance to the topic over time. Less related keywords are eliminated. Through this process, the remaining more relevant keywords are used to represent the topic.

Once keywords are extracted, then the networks are formed with each time period. The topic networks are created according to each topic keyword and time period. Every network was monitored to reveal how it evolves and when one topic cluster is absorbed by another, defined as *cannibalism*. After networks are examined visually, network properties are calculated to explain when the network shows cannibalism. Basic network properties like network size, graph density, and modularity are calculated.

3.1. Research publication activity

Research publication activity over time can be considered a representation of knowledge. It is analyzed by the number of publications on each topic and the relation between topics. The activity on a specific topic can be viewed by the number of publications on a specific topic analyzed by keyword occurrences. The keyword set used to define each publication can be supplied by the author or the publication journal based on a predefined set of keywords or extracted from the title or the abstract. The basic time frame for evaluating topic publication activity was set to one year. Smaller time frames were analyzed but seemed less significant due to the timeline of the research activity, which is periodic over the course of the year so smaller time segments cannot be used.

Research publications in different domains were analyzed. The research topics in each specific time period were identified. There are several clusters in the topic. Many keywords are included in a cluster. A keyword in a topic is a sub-topic or a related topic of the research area. Cannibalism in the network in this study means one research topic keyword cluster is consumed into other clusters. It indicates that keywords in the different clusters are getting close to each other.

3.2. Related topic identification

Evaluating connectivity, or communication, between research topics is based on identifying first the related topics. This is done by classifying multiple topics that appear in the same research article, as identified by the selected keywords. This means that if an article contains any two topics, then those two topics can be considered to be related to each other. Once two topics are marked as related, the change of the topics' activity over the whole time period viewed was analyzed.

The research topic analysis method includes the following steps:

Fig. 1. Example of relevant keyword pairs and occurrences.

- Select a topic and download articles parsed by fields within the time period (field parsing includes title, author keywords, journal assigned keywords, abstract, and article content).
- Extract the publication year and list of specific topic keywords associated with each article.
- Count number of appearances per year of each keyword.
- Identify possible relations between keywords based on multiple keywords appearing in a single article.
- Identify changing trends over time between possible relevant keywords:

For every two related keywords

For every year
If both keywords have at least two years ascending values at similar times
If both keywords have descending values in the following year
Mark keyword trend similarity as “relevant”
Else
Mark keyword trend similarity as “irrelevant”
Else
Mark keyword trend similarity as “irrelevant”

Fig. 1 displays identified possible similar keyword pairs with their temporal occurrences. Examples of keywords identified as “relevant” according to above method are marked in red boxes.

3.3. Evaluating performance of keywords extraction

Only keywords marked as relevant were analyzed in this research. Relevant represents both the article author’s opinion and occurrences in a specific time period.

Relevant keyword terms $T1, T2$ for a given set of articles $\{A_1, \ldots, A_n\}$, time series $\{t_1, \ldots, t_n\}$, where $|T1_{t_j}|$ is the number of $T1$ term appearing at time t_j are defined as:

$$\left\{ (T1,T2) |\ T1,T2 \in A_i,\ \exists j,\ \left|T1_{t_j}\right| > \left|T1_{t_{j-1}}\right|, \left|T1_{t_{j-2}}\right|, \left|T1_{t_{j+1}}\right| \right.$$
$$\left. \&\left|T2_{t_j}\right| > \left|T2_{t_{j-1}}\right|, \left|T2_{t_{j-2}}\right|, \left|T2_{t_{j+1}}\right| \right\}$$

The two keyword terms $T1$ and $T2$ appeared in the same article and received local maximum of number of appearances in articles published in the same year.

Other methods of identifying relevant keywords such as identifying correlation were considered. However, methods such as Pearson Correlation assume normal distribution and linearity of the data which does not exist since the data is represented by sparse multiple peaks of keyword occurrences followed by long empty time periods of no activity on the topic. Other methods, such as Spearman Correlation, require the variable to be monotonically related to the other variable, whereas our data is non-monotonic. Many methods exist which try to identify the degree of correlation between two variables. However, we consider keywords to be relevant to a specific topic if there is a single change (ascending followed by descending) of the occurrence of two keywords at the same time and at least one author is considered to identify both keywords as relevant in the same article.

The ratio of related keywords and eliminated keywords was defined as the number of relevant keywords over the number of all terms (sum of number of relevant and irrelevant). The ratio of relevant keywords and eliminated keywords differs by topic and period.

Table 1 shows a comparison of other methods correlation categorization compared to the cannibalism analysis method used here. The table shows that other methods require the data to be monotonic in the case of Spearman correlation or normal and linear distribution in case of Pearson correlation. Furthermore, if Spearman correlation can be viewed as requiring perfect correlation and the Pearson correlation as requiring imperfect correlation, the proposed method only requires local correlation in a limited time period. In these categories the method used here is least restrictive compared to commonly used methods. The method used here allows the analysis of data which is more heterogeneous compared to the standard correlation methods. In addition, the method has an advantage over the statistical methods by including high leverage points and therefore taking into account outlier points. These outlier points depict new research, and the present work focuses on the change over time of these new research topics. The goal of this research is to be able to classify topics that are not already identified as correlated and to identify how new clusters are formed over time based on existing clusters.

The ratio is different by parsed field even in the same topic. There are fields to describe data elements of articles predefined by each database. Generally an article contains author-defined keywords and journal-defined keywords. As a result, the ratio can be different according to which field is used to extract keywords. In this research, networks consist of keywords as nodes, so using the best field to extract keywords is important. The analysis of the relevance of each field is described in the Experiments Section.

3.4. Network creation

Medical topic networks are created with extracted keywords after eliminating less relevant ones. The data listed in Fig. 1 includes keywords, publication years, and the number of co-appearances between two keywords. From this data set the topic network can be designed using a force-directed graph algorithm. The purpose of the force-directed graph drawing algorithm is to position the nodes of a graph in two-dimensional or three-dimensional space so that all the edges are of more or less equal length and there are as few crossing edges as possible, by assigning forces among the set of edges and the set of nodes, based on their relative positions, and then using these forces either to simulate the motion of the edges and nodes or to minimize their energy [39].

Force Atlas is a force-directed layout close to other algorithms used for network spatialization [40]. Forces in a force-directed layout make nodes repel each other while edges attract their connected nodes. These forces let the movements converge to a stable state. The repulsion force F is proportional to the product of the

Table 1
Correlation methods categorization comparison.

Method	Data relation requirements	Probability density	Probability distribution	Correlation requirements	Data consistency	Leverage points
Pearson	Non-monotonic	Normal	Linear	Imperfect	Homogeneity	Low
Spearman	Monotonic	Normal	Non-linear	Perfect	Homogeneity	Low
Cannibalism	Non-monotonic	None	Non-linear	Local	Heterogeneity	High

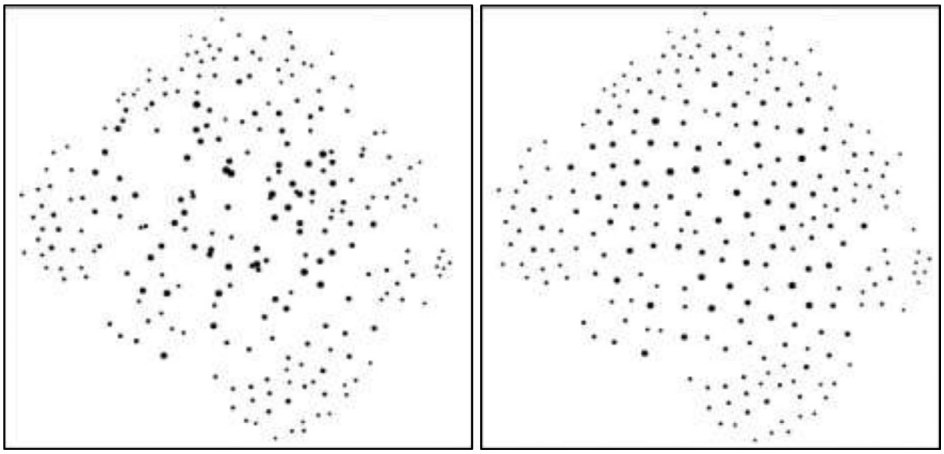

Fig. 2. Layout algorithm with and without nodes overlapped.

degrees of two nodes plus one $(deg + 1)$. This value changes by movement speed, scale, gravity, and other factors for the layout. The coefficient k is defined by the user settings.

The force between two nodes $F(n1, n2)$ is defined as:

$$F(n1, n2) = k\frac{(\deg(n1) + 1)(\deg(n2) + 1)}{d(n1, n2)}$$

The denominator $d(n1, n2)$ means distance between two nodes. This formula is a modified version of an existing formula for clustering energy models. Noack [41] suggested a similar formula without $(deg + 1)$ but just with the deg. The degree plus one layout algorithm is important to cover a node with zero degrees but also have some repulsion force.

Preventing node overlap is also important for network visualization. Fig. 2 shows an overlapping case and preventing an overlapping case using the Gephi network analysis and visualization software [42], with k value set to 100. The graph on the left in Fig. 2 is an original graph without links of the topic of mycology from 1985 to 1994. The graph on the right is the same graph with the prevention of overlapping nodes. If the distance between two nodes is positive, then there is no overlap. In this case, $d(n1, n2)$ is replaced by $d'(n1, n2) = d(n1, n2) - size(n1) - size(n2)$ to compute the force:

$$F'(n1, n2) = k\frac{(\deg(n1) + 1)(\deg(n2) + 1)}{d'(n1, n2)}$$

If the distance between two nodes is negative, then the two nodes are overlapped. Preventing node overlap is also important for network visualization. Then there is no attraction between them and the repulse force is calculated as follows:

$$F'(n1, n2) = k'(\deg(n1) + 1)(\deg(n2) + 1)$$

The last case: if $d(n1, n2) = 0$, then there is neither attraction nor repulsion force.

After the layout algorithm ran to form a network, nodes are colored by their cluster and the size represents its degree. The size of labels is proportional to node size.

Fig. 3 is an example of a created network on the topic of microbiology.

3.5. Network measures

The distance between a pair of nodes in a graph is the length of the shortest path between the two nodes. This is the definition of *network diameter* [1]. By definition, network diameter can be computed after all the shortest paths of every pair of nodes are calculated. A network consists of nodes and edges. *Degree* is the number of edges linked to a node. Every node has degree of 0 or a positive value. If a node has 0 degrees, then this node is isolated. There are no other nodes connected to this node. A higher degree means there are many directly connected nodes. The average degree of a network is computed as twice the number of edges divided by the number of nodes.

The *density of a network* is defined as twice the number of edges divided by the number of nodes (N) multiplied by $(N - 1)$. The denominator $N(N - 1)$ means the number of possible edges. It is an indicator for the general level of connectedness of the graph [43]. *Modularity* is a measure of the quality of a particular division of a network [44]. This value is larger than or equal to –1/2 and smaller than 1. If the modularity value is high, then the graph has a greater chance to split into two communities [45]. In this study, modularity is used for analyzing whether nodes are getting merged. In other words, the meaning of clusters in networks has weakened over time.

3.6. Cannibalism identification

Created networks of a topic were analyzed by network measures. For each topic, seven different networks are generated and analyzed by each time period. For each topic four networks represent a sliding window over time: 1975–1984, 1985–1994, 1995–2004, and 2005–2014. Another four networks represent a cumulative shrinking time window: 1975–2014, 1985–2014, 1995–2014, and 2005–2014. The last network is identical in both views. The sliding time window allows only a narrow time view of the activity in the research topic while the shrinking window allows a broader viewpoint of the networks with a cumulative time period. If there is any tendency for one cluster to be consumed by another cluster, then the topic is considered a cannibalism topic. If

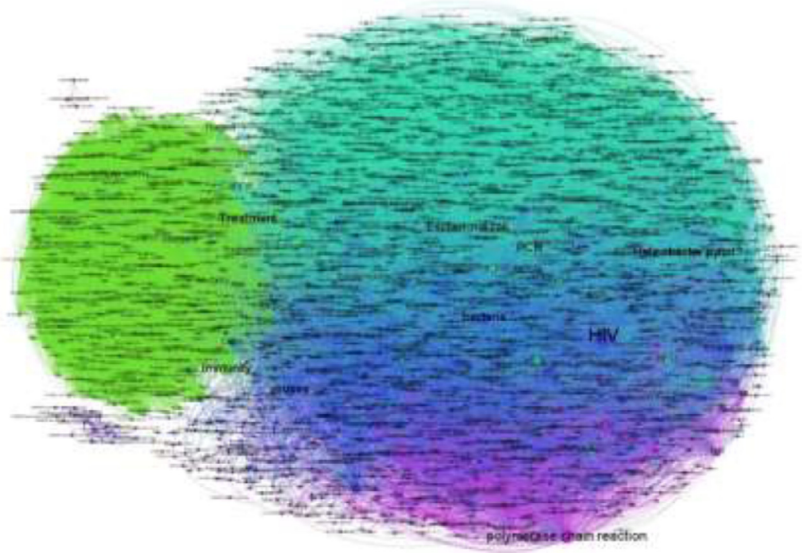

Fig. 3. An example network of microbiology.

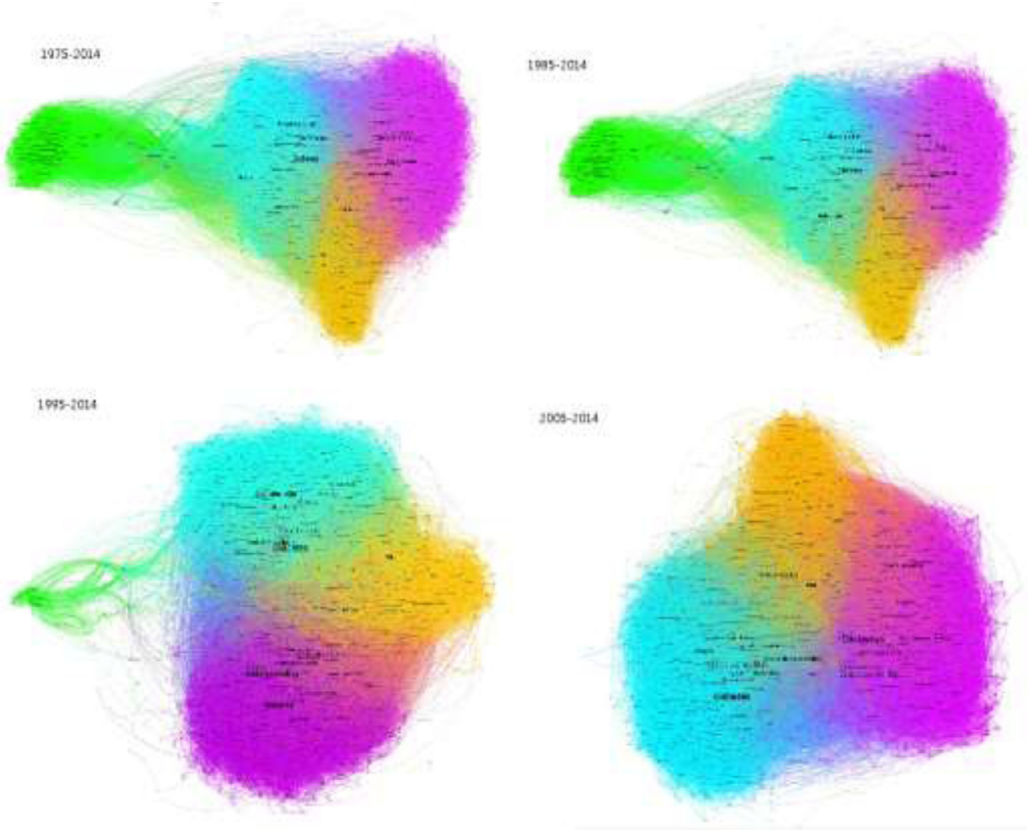

Fig. 4. A cannibalism topic - diabetes.

one cluster is not clearly absorbed by another cluster but the overall nodes are merged together and form one large cluster, then it is considered a weak cannibalism topic. Weak cannibalism is defined as two clusters having each at least 30% of the overall nodes in a given time period followed by merger of the two clusters in the following time periods. Cannibalism identifies one topic consuming the other while weak cannibalism represents a merger between the two topics where it is not clear which cluster absorbed the other. If there are no changes of the clusters, then the topic is non-cannibalism.

Figs. 4–6 illustrate each topic classification. Fig. 4 shows how a diabetes network changes over time. One separate cluster and the other clusters are combined together. Fig. 5 is an example of a weak cannibalism topic in the epidemiology network. In Fig. 5 the first two time periods display two clusters that are set apart with each containing more than 30% of the nodes. The two clusters form together in the last period. There is no obvious cannibalism pattern, but the change of clusters makes all the nodes form one large component. An example of a non-cannibalism topic is shown in Fig. 6. No change of clusters is observed in the oncology network.

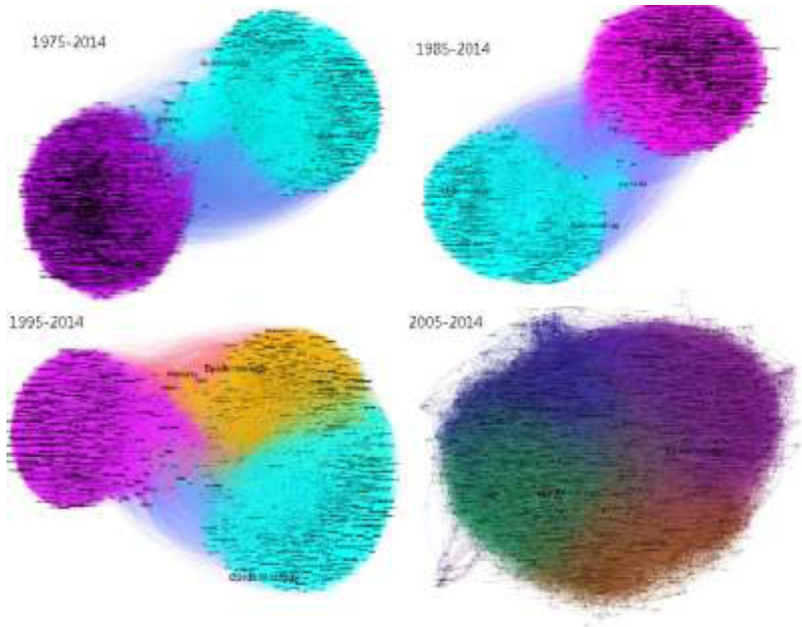

Fig. 5. A weak cannibalism topic – epidemiology.

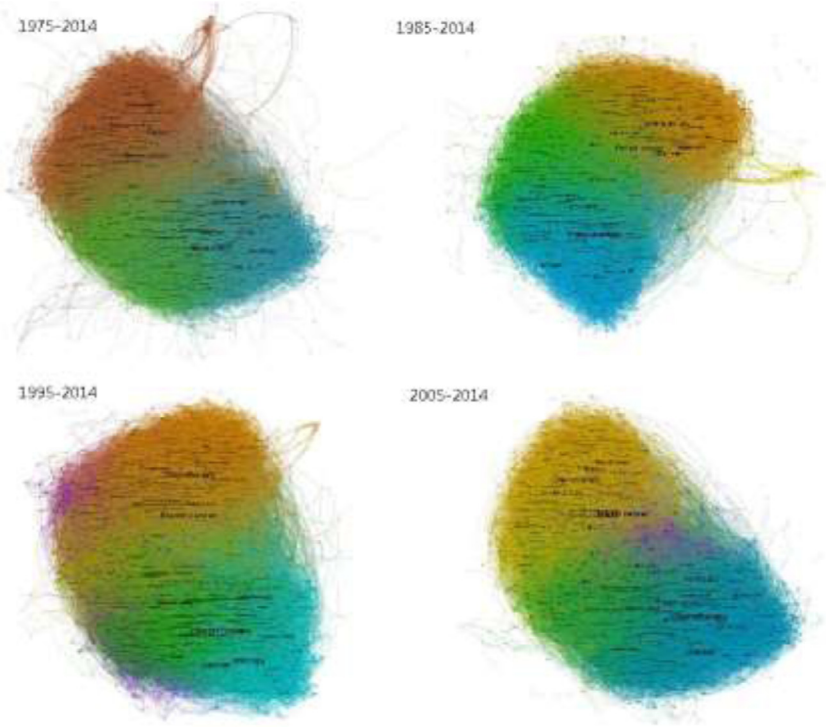

Fig. 6. A non-cannibalism topic - oncology.

4. Experiments

4.1. Data

PubMed and Web of Science provide access to multiple databases of references and abstracts on biomedical topics. Articles were extracted to analyze research publication activity from both data sources according to topic keywords such as: *angina, diabetes, diphtheria, epidemiology, genetics, hematology, hepatitis, immunology, infectious disease, in vitro fertilization, microbiology, mycology, nephrology, obesity, obstetrics, oncology, ophthalmology, orthopedic, poliomyelitis,* and *virology.*

The analysis time span is 40 years, from 1975 to 2014. Related keywords were identified based on one year time periods consolidated into ten year time windows. For each topic four networks represent a sliding window over time: 1975–1984, 1985–1994, 1995–2004, and 2005–2014. Another four networks represent a shrinking time window: 1975–2014, 1985–2014, 1995–2014, and 2005–2014. Some topics, including angina, hematology, nephrology, ophthalmology, orthopedic, and virology, did not have

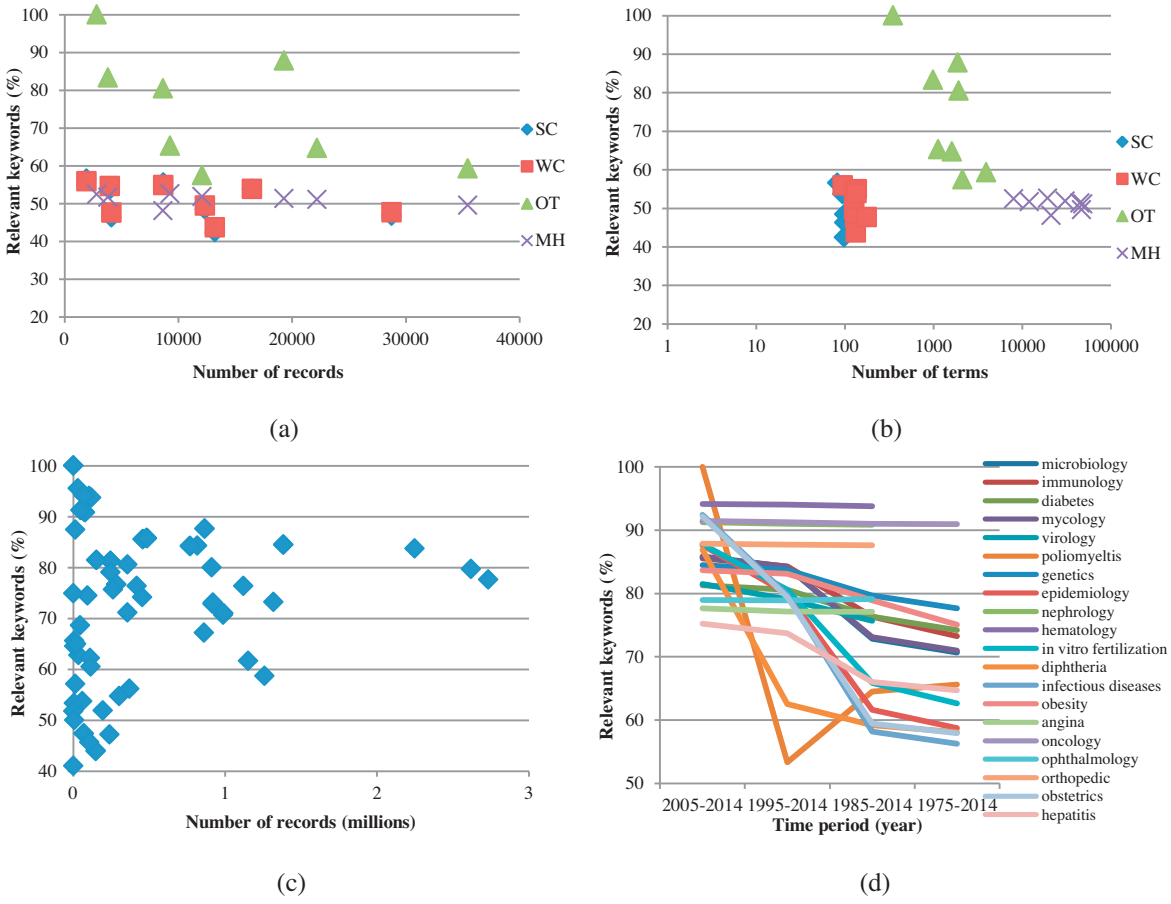

Fig. 7. Keyword field tag - relevant keywords vs. number of records (a) and number of terms (b) Relevant keywords vs. number of records (c) and time period (d).

sufficient articles to create a network from 1975 to 1984. The number of records for each topic ranged from 8,845 (poliomyelitis) to 2,734,571 (genetics).

Two keyword search field tags were used from each data source. Web of Science has WC (Web of Science Category) and SC (Subject Category). PubMed has MH (MeSH Terms) and OT (Other Term). MeSH (Medical Subject Headings) is a vocabulary thesaurus for indexing and organizing terms in the life sciences, developed and controlled by the United States National Library of Medicine (NLM). MH and WC are journal or web database assigned terms in contrast to OT and SC, which are author-assigned tags.

4.2. Experiments

Extracted keywords are different by tags and source of data, so an experiment to select the best tag was performed. To compare the performance of each tag, the ratio of relevant keywords and irrelevant keywords was computed.

The occurrences of a keyword of each year were analyzed to compare the trend similarity of every pair of keywords. If two keywords have a similar trend, then the keywords are marked as relevant and are connected in the network. On the other hand, if two keywords have no similar trend, then they are marked as irrelevant and there is no link between them.

Only trend similarity of keywords marked relevant is included in the networks. As a result, a total 128 networks of 20 topics with 7 different periods are created. The networks are analyzed to identify cannibalism. The topics are divided into three categories: cannibalism, weak cannibalism, and non-cannibalism, described in Section 3.6.

4.3. Results

4.3.1. Keyword classification

Fig. 7(a) analyzes the type of keyword field tags in classifying relevant keywords versus number of records. The highest is when extracting by using the OT field tag. The most relevant keywords are extracted with OT compared to the other three tags in every topic regardless of the number of records. The extraction of relevant keywords is about 50% using WC, SC, and MH tags, while for every value using OT it is over 50%.

Fig. 7(b) also shows the percent of relevant keywords compared to number of terms. The X-axis is in logarithmic scale. MH and OT tags from PubMed extract many more terms compared to Web of Science data. Although the MH tag can extract the most terms, the percentage of relevant keywords is the highest using the OT tag. Fig. 7(a) and (b) show that OT is the best performing tag. As a result, it was used for the keyword network analysis in the following results.

Fig. 7(c) shows that the percent of relevant keywords is not affected by the number of records. Fig. 7(d) illustrates how the percent of relevant keywords changes over time in 20 topics. The graphs are decreasing to the right side. Therefore, the keywords are less related to each other in the same topic as the time window expands.

4.3.2. Topic categorization

Half of the twenty topics show cannibalism and six topics have weak cannibalism. The other four topics are categorized as non-cannibalism topics. Table 2 shows each category type and relevant topics.

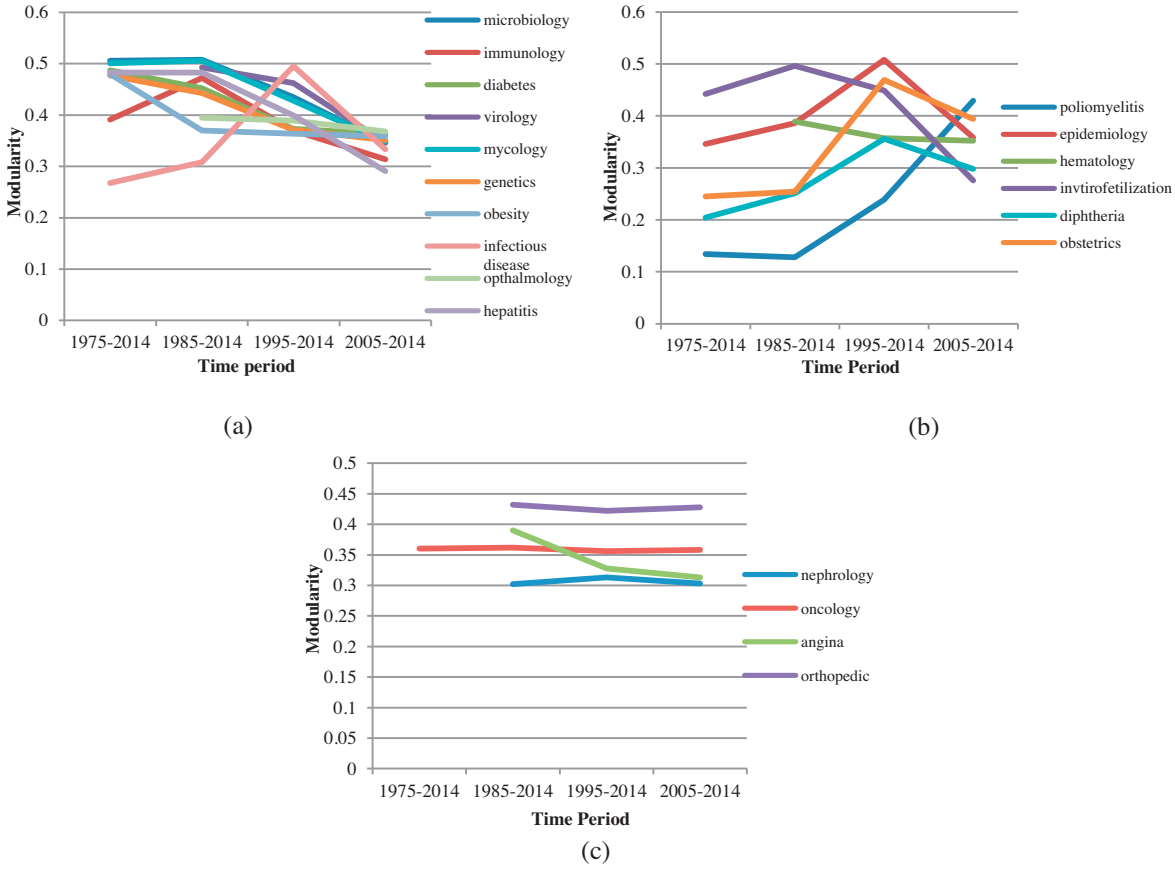

Fig. 8. Modularity of topic cannibalism (a), weak cannibalism (b), and non-cannibalism (c).

Table 2
Topic categorization.

Type	Topic
Cannibalism	Diabetes, Genetics, Hepatitis, Immunology, Infectious disease, Microbiology, Mycology, Obesity, Ophthalmology, Virology
Weak cannibalism	Diphtheria, Epidemiology, Hematology, In vitro fertilization, Obstetrics, Poliomyelitis,
Non-cannibalism	Angina, Nephrology, Oncology, Orthopedic

4.3.3. Network measures

The most distinct network property between cannibalism and non-cannibalism topics is changes of modularity values. Modularity is one measure for the partitioning of a network [46]. Fig. 8(a) shows the values are decreasing as the periods are getting shorter. This change concurs with the cannibalism; nodes in one cluster are consumed by another cluster. In general, the values decrease in Fig. 8(b) when the period is changed to the past ten years for weak cannibalism topics. However, Fig. 8(c) shows the modularity values of non-cannibalism topics do not change much.

Another network characteristic is average degree of the networks. As explained in Section 3.5, all of the nodes in a network have degree, which represents how many links are connected to the node and the average degree of a network is calculated by twice the number of edges divided by the number of nodes in the network. If the average degree is high in a network, then there are many connections among the nodes. The X-axis of Figs. 9 and 10 represents the time periods of networks and the Y-axis represents average degrees.

Fig. 9(a) shows that the average degree of cannibalism topics changes slightly. In Fig. 9(b), the values of weak cannibalism topics decrease rapidly. Fig. 9(c) shows there are no obvious changes in non-cannibalism topic networks.

On the other hand, Fig. 10 shows that the networks values change considerably regardless of type of topics in the ten-year sliding time window. Fig. 10(a) shows the values go up and down at every period of cannibalism topics. Fig. 10(b) shows the average degree values of weak cannibalism topics peaked at 1985 to 1994 and then decreased. The values of non-cannibalism topics change opposite to those of weak cannibalism topics in Fig. 10(c). They are boosted in the last period.

5. Discussion

The experiments show that there is cannibalism in the network of medical topics, where cannibalism of a network is defined as one cluster consumed into other clusters. The results identified that cannibalism occurs in 50% of topics analyzed in medicine and classified the main reasons that lead to cannibalism. Some of the topics do not exhibit exact cannibalism, but nodes get closer to each other. Overall clusters move to form one large component.

20 topics were selected in the medical domain to analyze the temporal change of the networks. The medical topics were analyzed because there is a well-defined structure for the research articles of the area of medicine. Although the twenty topics cannot represent all the topics in the medical domain, this research analyzed a large data set. In total, 9,730,671 articles were analyzed. The size of topics varies from 8,845 articles of poliomyelitis to 2,734,571 articles of genetics. In addition, 1,017,486 terms are extracted from the research publications. The number of terms is less

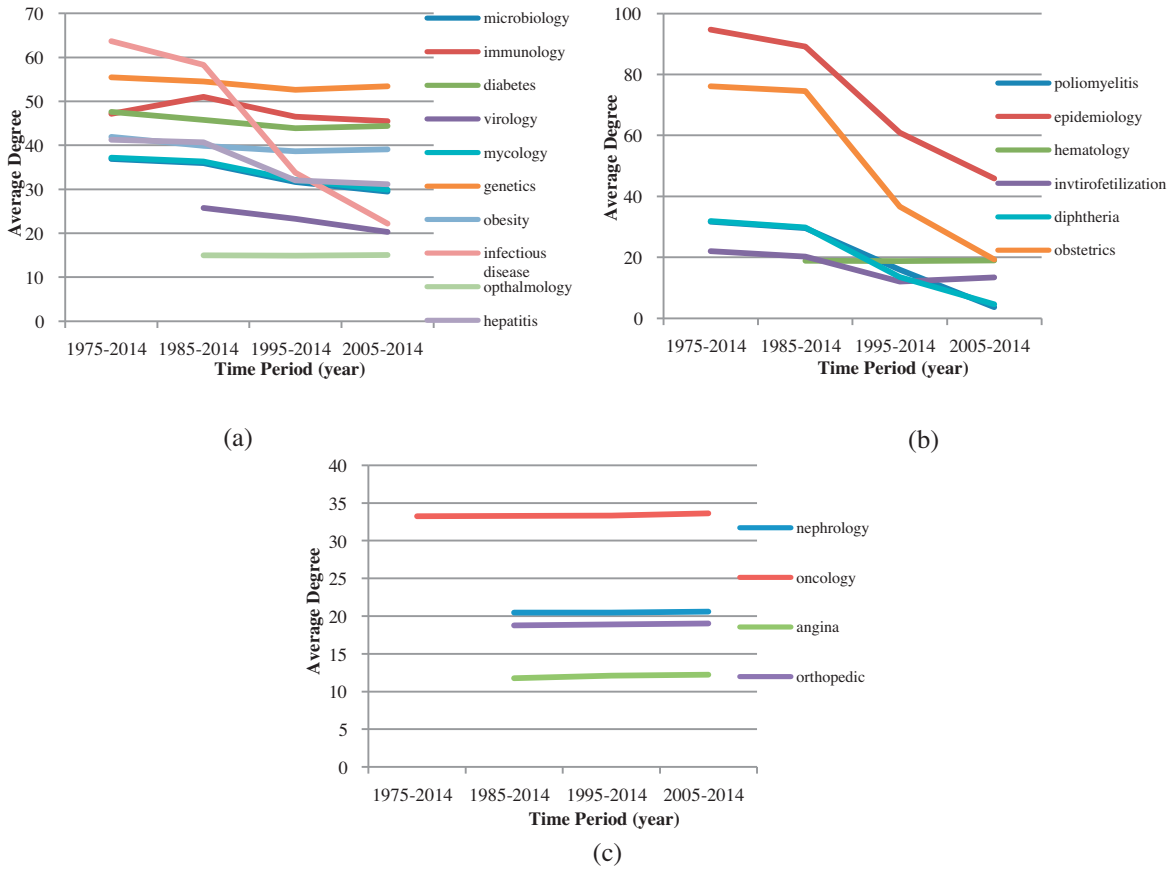

Fig. 9. Average degree (cumulative period) of topic cannibalism (a), weak cannibalism (b), and non-cannibalism (c).

than the number of articles because many keywords overlapped in a same topic and between topics.

The medical topic networks consist of nodes and edges. Nodes are keywords in the topic and edges are co-occurrence between two keywords in the same article. The keywords of an article are defined by author or journal. There are two tags to indicate keywords of each database: PubMed and Web of Science. The experiments showed that PubMed data contain many more relevant terms compared to Web of Science data, which could be contributed to the PubMed specialization in medical articles. Additionally, the OT tag was the best to extract keywords.

The twenty topics were classified into three categories: cannibalism, weak cannibalism, and non-cannibalism. The results indicated that ten topics have the cannibalism pattern. There was no obvious cannibalism, but clusters are combined together over time in six weak cannibalism topics. The remaining four topics are categorized as non-cannibalism topics. There was no noticeable cluster movement in the non-cannibalism topics.

One limitation of this research is the use of trend similarity to detect the relation of two keywords, as described in Section 3.2. Trend similarity is a method to compare the occurrences of two keywords. Only links between similar trend keywords were included in the networks. However, the relation between them can be changed to other measures. Existing mathematical correlation measures like Pearson's product-moment coefficient could show other results in detecting relations between them.

One issue that should be considered is differentiating between keyword matching and concept matching, as stated in Cambria and White [47]. In Section 3.1 we mention that the keyword set used to define each publication can be supplied by the author or the publication journal based on a predefined set of keywords or extracted from the title or the abstract. However, to avoid keywords that have multiple meanings, we used in the experiments only concepts supplied by the authors or by the journal. These concepts can include single keyword or multiple keywords.

Another limitation of this research is the lack of diverse approach to analyze the networks. The network measures were recorded for each topic and period to distinguish between cannibalism and non-cannibalism topics. But as explained in Section 4.3, only modularity and average degree were valid to identify each topic type. Other supervised learning methods such as decision trees, neural networks, or ensemble methods can be applied to analyze the networks and identify when the cannibalism occurs.

Despite the limitations, this research showed that there is a similar pattern of cluster changing of many topics. Like scientists collaborate with other researchers, even from different fields, the research topics also collaborate with each other. Moreover, nowadays as multidisciplinary fields have emerged, the boundary of clusters in the medical topics has weakened.

6. Conclusion

The paper presents cannibalism in medical topic networks. Cannibalism of a network means one cluster is consumed into other clusters. Twenty topics were analyzed for the moving of nodes or clusters of the networks. Research publication activities of each topic were collected and representative terms were selected to create networks of the topics. The topics were categorized by existence of cannibalism. Ten out of twenty topics have obvious cannibalism in their networks. Six topics do not show cannibalism clearly, but the clusters of the networks are also converged. Overall, the distances between clusters are being shortened.

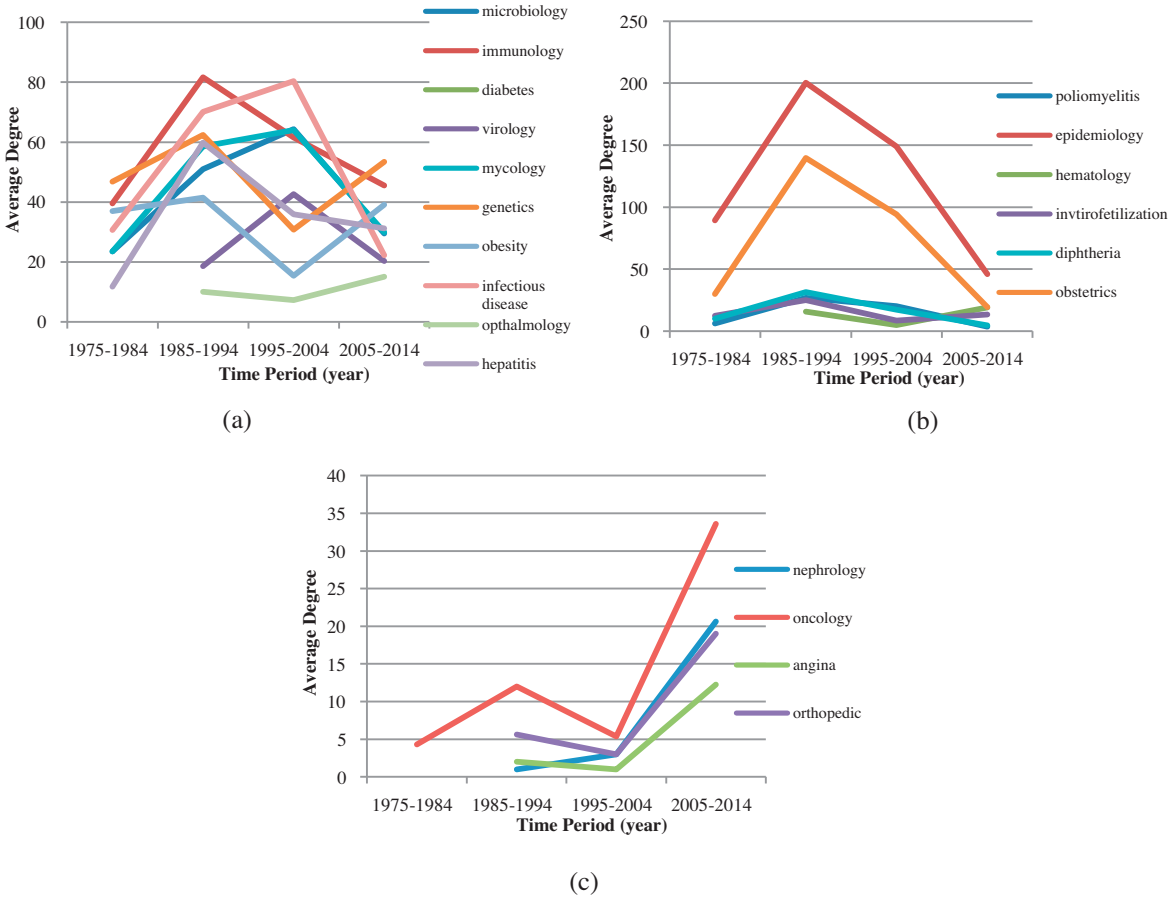

Fig. 10. Average degree (sliding window period) of topic cannibalism (a), weak cannibalism (b), and non-cannibalism (c).

For future study, a greater selection of topics of different domains can be analyzed. Future research can expand to other domains of medical topics, and domains in other fields, including science, technology, and humanities, can be analyzed. In addition, the meaning of keywords of topics can be analyzed to detect more clearly the emergence of core terms or categories.

Acknowledgment

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. B0101-16-1272, Development of Device Collaborative Giga-Level Smart Cloudlet Technology).

References

[1] M. Newman, Who is the best connected scientist? A study of scientific coauthorship networks, Lect. Notes Phys. 650 (2004) 337–370.
[2] S. Wasserman, K. Faust, Social Network Analysis: Methods and Applications, Cambridge University Press, New York City, New York, U.S.A., 1994.
[3] J. Kim, C. Perez, Co-authorship network analysis in industrial ecology research community, J. Ind. Ecol. 19 (2) (2015) 222–235.
[4] M. Newman, The structure of scientific collaboration networks, PNAS 98 (2) (2001) 404–409.
[5] S. Borgatti, D. Halgin, On network theory, Org. Sci. 22 (5) (2011) 1168–1181.
[6] R. Burt, Structural holes and good ideas, Amer. J. Sociol. 110 (2) (2004) 349–399.
[7] D. Obstfeld, Social networks, the tertius iungens orientation, and involvement in innovation, Admin. Sci. Quart. 50 (1) (2005) 100–130.
[8] R. Sparrowe, R. Liden, S. Wayne, M. Kraimer, Social networks and the performance of individuals and groups, Acad. Manage. J. 44 (2) (2001) 316–325.
[9] P. Anklam, Net Work: A Practical Guide to Creating and Sustaining Networks at Work and in the World, Heinemann, Woburn, 2007.
[10] R. Burt, Structural Holes: The social structure of competition, Cambridge: Harvard University Press, 1992.
[11] M. Killduf, D. Krackhardt, Bringing the individual back in: a structural analysis of the internal market for reputation in organizations, Acad. Manage. J. 37 (1) (1994) 87–108.
[12] D. Brass, K. Butterfield, B. Skaggs, Relationships and unethical behavior: a social network perspective, Acad. Manage. Rev. 23 (1) (1998) 14–31.
[13] K. Heyman, Making connections, Science 5787 (604-606) (2006) 313.
[14] R. Ferrer, C. Janssen, R. Sole, The topology of technology graphs: small world patterns, Phys. Rev. E 64 (2001) 046119.
[15] G. Nan, C. Zang, R. Dou, M. Li, Pricing and resource allocation for multimedia social network in cloud environments, Knowl. Based Syst. 88 (2015) 1–11.
[16] J. Montoya, S. Pumm, R. Sole, Ecological networks and their fragility, Nature 442 (2006) 259–264.
[17] K. Goh, M. Cusick, D. Valle, B. Childs, M. Vidal, A. Barabassi, The human disease network, PNAS 104 (21) (2007) 8685–8690.
[18] F. Cheong, B. Corbitt, A social network analysis of the co-authorship network of the Pacific Asia conference on information systems from 1993 to 2008, in: Proceedings of Pacific Asia Conference on Information Systems (PACIS 2009), Hyderabad, India, 2009.
[19] G. Davis, H. Greve, Corporate elite networks and governance changes in the 1980 s, Am. J. Sociol. 103 (1) (1997) 1–37.
[20] S. Borgatti, P. Foster, The network paradigm in organizational research: a review and typology, J. Manage. 29 (6) (2003) 990–1013.
[21] T. Stuart, Network positions and propensities to collaborate: an investigation of strategic alliance formation in a high-technology industry, Admin. Sci. Quart. 43 (3) (1998) 668–698.
[22] L. Garton, C. Haythornthwaite, B. Wellman, Studying online social networks, J. Comput. Mediat. Commun. 3 (1) (1997).
[23] S. Jung, A. Segev, Analyzing future communities in growing citation networks, Knowl. Based Syst. 69 (2014) 34–44.
[24] G. Melin, O. Persson, Studying research collaboration using co-authorships, Scientometrics 36 (3) (1996) 363–377.
[25] J. Osca-Lluch, E. Velasco, M. Lopez, J. Haba, Co-authorship and citation networks in Spanish history of science research, Scientometrics 80 (2) (2009) 373–383.
[26] N. Patel, Collaboration in the professional growth of American sociology, Soc. Sci. 6 (1973) 77–92.
[27] J. Travers, S. Milgram, An experimental study of the small world problem, Sociometry 32 (4) (1969) 425–443.
[28] B. Kademani, V. Kalyane, A. Kademani, Scientometric portrait of nobel laureate S.Chandrasekhar, Int. J. Scientometrics Informetrics 2 (2-3) (1996) 119–135.

[29] B. Kademani, V. Kalyane, S. Jange, Scientometric portrait of Nobel laureate Dorothy Crowfoot Hodgkin, Scientometrics 45 (2) (1999) 233–250.
[30] J. Kauffman, A. Kittas, L. Bennett, S. Tsoka, DyCoNet: a gephi plugin for community detection in dynamic complex networks, PLoS ONE 9 (7) (2014) 101357.
[31] A. Barabasi, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.
[32] A. Lancichinetti, S. Fortunato, F. Radicchi, Benchmark graphs for testing community detection algorithms, Phys. Rev. E 78 (4) (2008) 046110.
[33] M. Newman, Modularity and community structure in networks, PNAS 103 (23) (2006) 8577.
[34] V. Blondel, J. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. 2008 (10) (2008) 10008.
[35] G. Palla, A. Barabasi, T. Vicsek, Quantifying social group evolution, Nature 446 (2007) 664–667.
[36] P. Holme, J. Saramaki, Temporal networks, Phys. Rep. 519 (2012) 97–125.
[37] W. Quattrociocchi, F. Amblard, E. Galeota, Selection in scientific networks, Soc. Netw. Anal. Mining 2 (3) (2011) 229–237.
[38] J. Michel, Y. Shen, A. Aiden, A. Veres, M. Gray, T.G.B. Team, J. Pickett, D. Hoiberg, D. Clancy, P. Norving, J. Orwant, S. Pinker, M. Nowak, E. Aiden, Quantitative analysis of culture using millions of digitized books, Science 331 (2011) 176–182.
[39] S.G. Kobourov, Spring embedders and force-directed graph drawing algorithms, arXiv 1201.3011 (2012).
[40] M. Jacomy, T. Venturini, S. Heymann, M. Bastian, ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software, PLoS ONE 9 (6) (2014) e98679.
[41] A. Noack, Energy models for graph clustering, J. Graph Algorithms Appl. 11 (2) (2007) 453–480.
[42] M. Bastiam, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks, Proceeding of the third international AAAI Conference on Webblogs and Social Media, 2014.
[43] E. Otte, R. Rousseau, Social network analysis: a powerful strategy, also for the information sciences, J. Inf. Sci. 28 (6) (2002) 441–453.
[44] M. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 026113.
[45] L. Danon, A. Daz-Guilera, J. Duch, A. Arenas, Comparing community structure identification, J. Stat. Mech. Theory Exp. 2005 (9) (2005) 09008.
[46] S. Fortunato, M. Barthelemy, Resolution limit in community detection, PNAS 104 (1) (2007) 36–41.
[47] E. Cambria, B. White, Jumping NLP curves: a review of natural language processing research, IEEE Comput. Intell. 9 (2) (2014) 48–57.