Journal of Informetrics 16 (2022) 101242

Contents lists available at ScienceDirect

Journal of Informetrics

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

Exploring scientific trajectories of a large-scale dataset using topic-integrated path extraction

Erin H.J. Kim [a], Yoo Kyung Jeong [b], YongHwan Kim [c], Min Song [d,*]

[a] Department of Library and Information Science Education, College of Education, Kongju National University, Gongju 32588, Republic of Korea
[b] Department of Library and Information Science, Hannam University, Daejeon 34430, Republic of Korea
[c] Department of Library and Information Science, Cheongju University, Cheongju 28503, Republic of Korea
[d] Department of Library and Information Science, Yonsei University, Seoul 03722, Republic of Korea

ARTICLE INFO

Keywords:
Citation analysis
Healthcare informatics
Longest path
Main path analysis
Topic modeling

ABSTRACT

Main path analysis (MPA) is the most widely accepted approach to tracing knowledge transfer in a research field. In this study, we extracted multiple longest paths from the multidisciplinary academic field's citation network and integrating topic modeling to the extracted paths. We consider three main aspects of trajectory analysis when analyzing the represented documents through the extracted paths: emergence, authority, and topic dynamics. For path extraction, we adopt the longest path algorithm that consists of the following three steps: 1) topological sort, 2) edge relaxation, and 3) multiple path extraction. For topic integration into multiple paths, we employ latent Dirichlet allocation (LDA) by utilizing the topic-document matrix that LDA derives to select an article's topic from the citation network, where each article is labeled with the topic that is assigned with the highest topical probability for that article. We conduct a series of experiments to examine the results on a dataset from the field of healthcare informatics that PubMed provides.

1. Introduction

Understanding the lifecycle of science in any field is immensely important for scientists and practitioners. However, such a comprehensive understanding of knowledge diffusion requires a significant amount of time and effort. One simple and alternative approach to manual study is to search for the most highly cited papers in a field. This is useful because a highly cited paper is regarded to have relatively more importance or influence in a field (Garfield, 1979). This method provides a micro-level understanding in a field, meaning readers do not see the propagation of knowledge that the paper conveys in chronological order. A macro-level investigation searches for the mainstream in a particular field. Main path analysis (MPA), which is one popular approach to finding the mainstream in a specific discipline, can extract the relatively comprehensive core literature that scientists can understand without considerable time investment (Lu & Liu, 2013; Mina et al., 2007). This is because MPA reduces a citation network's complexity (Liu & Lu, 2012) and identifies the major flows of knowledge in a particular area (Verspagen, 2007) based on how scientific knowledge is disseminated from cited papers to citing papers.

However, current MPA techniques encounter challenges when applied to interdisciplinary or multidisciplinary studies in diverse academic fields that have developed more dynamic changes in subjects. A defining characteristic of a traditional main path method is that it significantly abstracts a complicated citation network; however, this benefit may not sufficiently capture the detailed knowledge transfer patterns of a mature, multidisciplinary field such as healthcare informatics, which is characterized by an immense number of

* Corresponding author.
E-mail addresses: erin.hj.kim@kongju.ac.kr (E.H.J. Kim), yk.jeong@hnu.kr (Y.K. Jeong), kimyonghwan@cju.ac.kr (Y. Kim), min.song@yonsei.ac.kr (M. Song).

https://doi.org/10.1016/j.joi.2021.101242
Received 4 January 2021; Received in revised form 10 October 2021; Accepted 8 December 2021
Available online 22 December 2021

publications and a wide span of history in the real world. In this case, longer knowledge transfer routes in a field may offer a better understanding of the knowledge transfer pathways. Furthermore, traditional MPA algorithms do not provide information about the topics of the individual nodes on a path. The research topology gained by MPA may not fully capture the topical connections among the citations because another work may be cited for a variety of reasons.To solve these problems, we propose a novel method called topic-integrated path analysis that combines longest paths and topic modeling to provide a better understanding of knowledge development in a multidisciplinary scientific field. To this end, we construct the weighted citation network using social network analysis and use the longest path algorithm to extract the main paths. We then integrate these main paths with topics derived through latent Dirichlet allocation (LDA) topic modeling (Blei, 2012), which uncovers the topics from an entire dataset. In our analysis, we investigate the following three aspects to understand knowledge diffusion in a research field: (1) academic emergence, (2) authority, and (3) topic dynamics. Emergence refers to the moment in time when knowledge originates. Based on the assumption that a discipline emerges from not only single knowledge, once there is a discipline's origin, convergent phenomena exist (Verspagen, 2007). Authority is at the point of occurring knowledge convergence and divergence (Batagelj, 2003; Hung et al., 2014). Knowledge convergence occurs because the paper cites many authoritative papers. Knowledge divergence occurs because many other papers cite the paper. Topic dynamics are observed in the knowledge flows that the topic presents (Kim Baek, & Song, 2018). By identifying the dominant topic of a paper, ongoing, changing, or active topics can we describe our methodology be detected over time.

The remainder of this paper is organized as follows: We discuss the related work for traditional approaches in the MPA section. In the section on the Proposed Approach, we describe our methodology, including data collection, measuring the path weight, extracting main paths, and integrating topics into the extracted main paths. For analyzing the experiment's results, we report the results of the extracted main paths and discovered topics from our exemplar dataset. In the Exploring Scientific Trajectories section, we discuss scientific trajectories in the context of the three aspects associated with the topic-integrated longest paths. In the conclusion section, we summarize the key points of our work and present future research directions.

2. Literature review

Hummon and Doreian (1989) conceptually proposed MPA algorithms using Eugene Garfield et al. (2003)'s historiography of the DNA field. They addressed three measures of "traversal counts" as edge weight in a citation network: node-pair projection count (NPPC), search-path link count (SPLC), and search-path node pairs (SPNP). The traversal count means the edge weight between a cited-citing pair that considers all possible paths through the network and converts a binary citation network into a weighted network. Then, they used a priority first search with three edge weights to generate main path. Batagelj (2003) proposed an efficient algorithm to utilize Hummon and Doreian's three edge weights for large-scale citation networks and proposed new edge weight, a search path count (SPC) algorithm. An SPC calculates the number of different paths (N) through a link (u, v), where v cites u, s denotes the start node, and t denotes the target node, as follows:

$$N(u, v) = n(s, u) \times n(v, t)$$

This equation multiplies the number of different s-u paths and v-t paths in a given network. Due to the less complicated computation required, the SPC method has been widely used in MPA research. However, some researchers argued for evaluation in weight assignments such SPLC (Liu et al., 2019) or SPNP (Kuan, 2020) in terms of knowledge diffusion.

Since the studies of Hummon and Doreian (1989) and Batagelj (2003), MPA has been exploited in searching the trajectory of disciplines such as medical science (Mina et al., 2007; Tu & Hsu, 2016), patents and technologies (Chen et al., 2020; Hung et al., 2014; Lathabai et al., 2018; Martinelli, 2012; Xiao et al., 2014; Yeo et al., 2016; Yu & Sheng, 2020), business and management (Calero-Medina & Noyons, 2008; Chuang et al., 2014; Henrique et al., 2019; Lu & Liu, 2013), environments (Epicoco et al., 2014), and education (Chen et al., 2019).

Lucio-Arias and Leydesdorff (2008) presented a longitudinal analysis of the 30 most-cited papers in the fields of "fullerene" and "nanotubes." They suggested using path-dependency to find the critical transition along the main path. The critical transition indicates a distribution shift in the knowledge chain. The critical transitions that they measured found two publication types, namely "forgotten" and "rewritten" publications. Liu and Lu (2012) proposed the alternative main paths, called the "global main path" and "key-route main path" to complement the original main path. The key-route refers to the largest traversal count in a given citation network. For a given number of key-routes, i.e., the first to second, or the first to tenth highest key-routes, the result may show up the multiple main paths. The key-route main path may overcome the problem, missing important edge, of both the local and global search methods. Hung et al. (2014) also adopted the multiple key-routes to examine the convergent and divergent patterns in the field of lithium iron phosphate battery. They found the distinct convergent and divergent pattern from the top 30 to the top 60 key-routes, then they extracted the main path at 60 links that resulted in 56 publications in the multiple main paths from the constructed citation network that contained 1480 publications. Martinelli and Nomaler (2014) proposed a "genetic approach" to identify technologically important patents in a patent citation network. In their method, they considered the backward mapping approach as a search scheme instead of the forward mapping approach of Hummon and Doreian's method and used a cut-off point for patent discovery to present multiple main paths.

Liu and Kuan (2016) considered decay factors for the traversal count in knowledge transfer. They called these factors arithmetic decay, geometric decay, and harmonic decay. In their results, the decay-parameterized main paths included more papers from the DNA literature citation network that Hummon and Doreian (1989) studied. Park and Magee (2017) proposed genetic backward-forward path analysis to find the main path in the patent citation network. Based on Martinelli & Nomaler (2014) concept of high

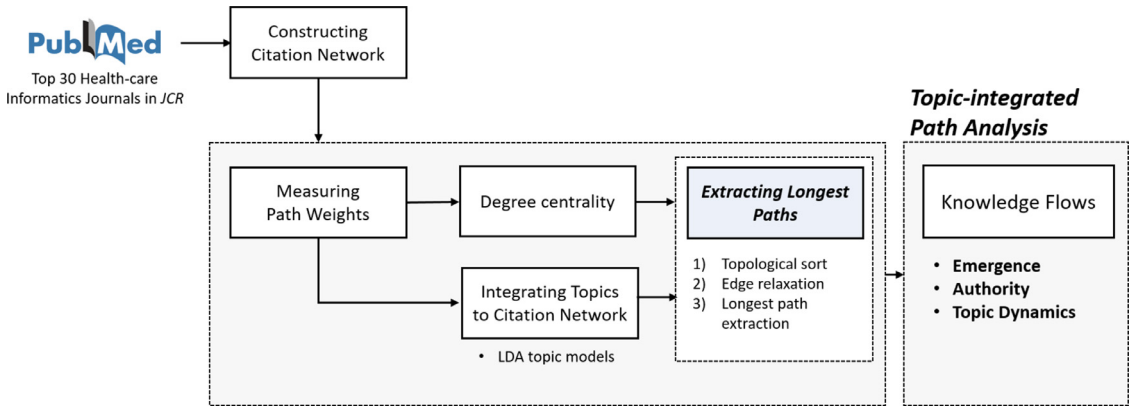


Fig. 1. The proposed approach's workflow.

persistence patents, they measured the knowledge persistence values of each node and selected some nodes having a persistence value over the threshold. Then, the backward- and forward-searching from the selected nodes identified the main paths.

Some studies attempted to analyze to identify topic diffusion along the main path. Tu and Hsu (2016) introduced the "conceptual development trajectory model" using key-route main path analysis with text mining. They assigned both document similarity and time to the articles in the main path and regarded the articles as a sort of topic. Kim et al. (2018) adopted the citation-influence model Dietz et al. (2007) proposed and measured the degree of the citation influence between citing and cited papers for presenting knowledge diffusion in the field of a p53 study. They focused on topic evolution along five critical paths that contained the normalized highest-influence path weights. In this study, we discover the topics of individual papers along with the extracted paths to investigate topic dynamics. Unlike (Kim et al., 2018), the results from our experiments contain the longest paths of papers over one from our dataset, allowing us to comprehend how topics emerge, change, merge, or split in the field of experimental study.

Direct citations occur as authors explicitly refer to related work in their own work; however, Evans et al. (2020) suggested using longest path analysis to map the sequence of knowledge transfer in a direct citation network because much of the information is disseminated not through explicit, direct citation but through a sequence of transfers across other papers. The longest path algorithm searches these paths exhaustively, enabling the identification of more sequences of knowledge transfer. In this study, we adopt longest path analysis to detect research trajectories in the healthcare field.

In addition, many of the preceding research data were limited to small-scale publications; thus, the resulting main paths were short in length. In this study, we attempted to map a large-scale citation network with a history of over four decades and extract various main path clusters that appear from the earliest to the latest year (Verspagen, 2007) by applying the longest path algorithm to the constructed citation network. Furthermore, we identified the topics of individual papers and integrated them into the extracted main paths to visualize topic dynamics in a multidisciplinary academic field.

3. Methodology: the proposed approach

In this section, we provide detailed descriptions of the proposed topic-integrated main path analysis. Fig. 1 illustrates the methodology and overall workflow of the topic-integrated longest path.

The proposed method combines two perspectives for path extraction. First, from the network perspective, we calculate edge weights between citing and cited papers to automatically extract multiple main paths from the weighted citation network. We employ the degree centrality measure to assign weights to the citation network and adopt the longest path algorithm for path extraction. Second, from the topic evolution or changing topics perspectives, the topic-integrated longest path reveals the dissemination of topics along the main paths. We present the extracted main paths combined with the identified topics of the individual components (i.e., papers contained in the main path) using topic modeling.

Another beneficial feature of our method is exhaustive path extraction. Traditional main path methods are outstanding for creating an outline of a complicated citation network; however, this may not capture the detailed knowledge transfer patterns in a large-scale citation network with a decades-long history. To overcome this problem, we exploit the longest path algorithm, which searches all possible paths exhaustively. The concept of the longest path algorithm is similar to global search that calculates the overall sum of the traversal weights (or counts) and identifies the overall significant paths in the network. However, the method of extracting the overall significant paths is different. The longest path sorts all nodes to be visiting and then searches every path from the root to all branches using a depth-first search method. The longest path considers all nodes for the next visit. In other words, the longest path combines local and global searches similar to the key-route.

In our experiments, we first perform data collection and configure weighted citation networks. We then extract multiple-clustered main paths using the longest path algorithm. Finally, we resolve the main path clusters including large components into topic-integrated main paths.

Following Batagelj's (2003) method to construct weighted citation networks, we first test the network's acyclicity and eliminate any loops. Second, we estimate the importance of the paths between citing-cited pairs. Additionally, we must investigate the MPA longitudinal analysis, which is a common practice for large-scale datasets (Brughmans, 2013). To satisfy this consideration, we collected 274,297 papers published over the past 46 years in a particular field. The following subsection details our methodology.

3.1. Data collection

For data collection, we chose healthcare informatics for our empirical study and adopted the top 30 journals in the healthcare informatics field based on JCR reports. We excluded the Health Affairs journal, which was ranked second in JCR in 2014, because of a lack of articles from PubMed. After selecting the top 30 journals, we collected seed articles from the 30 selected journal titles published, including authors, titles, abstracts, and publication years from 1970 to 2017. The seed data collection consisted of 89,369 papers and the number of articles collected from each journal. To construct a binary citation network, we collected citing papers from the seed papers using E-utilities, which is a program NCBI (https://www.ncbi.nlm.nih.gov/) provides. A total of 274,297 papers and 595,548 citing-cited pairs were included in our final dataset.

3.2. Removing cyclic data

Theoretically, a citation network must be acyclic. However, many citation networks are cyclic because of mutual citations, self-citation, or citation errors. In our study, we removed self-citation, mutual citations, and incorrect citations to create an acyclic network. If the PubMed ID (PMID) of a cited paper was equal to the PMID of a citing paper (i.e., if paper A cited paper A′ and paper A was identical to A′), that pair was removed. Additionally, we utilized self-developed Java programs to discard self-citations and citation errors. After refining the data, we were left with 595,060 citation pairs from the original dataset. There were 47,125 sources and 212,629 sinks for the empirical analysis in our study.

3.3. Measuring degree centrality and path weights

Since the citation network is a binary and unweighted network, we needed a weighting formula similar to SPC. The longest path algorithm finds the most significant path in a binary citation network, which constitutes the most exhaustive path in the network, but does not give any indication of how strongly the pairs are connected. To solve this problem, we employed social network analysis to measure computed edge weights for path weight (Elmacioglu & Lee, 2005). Social network analysis computes the importance and influence of the relationships between citing and cited pairs in a bidirectional way. We measured degree centrality as the number of links associated with a node. Degree consists of in-degree and out-degree in a directed network. We considered the importance of the node (paper) in the context of both cited and citing relations, so we used "degree," which identifies sums for in-degree and out-degree. A high value for in-degree indicates the document cites many other documents, whereas a high value for out-degree indicates a high frequency of citations by others in a citation network. Therefore, we computed the average of the in- and out-degrees of the citing-cited relationship nodes. We then applied the centrality mean of the two related nodes to the edge weights. For example, In Fig. 2(a), the value of degree centrality for N2 is 2 and the value for N3 is 4. Therefore, the edge weight between N2 and N3 is 3, which is the average degree centrality of both nodes N2 and N3.

3.4. Path extraction in order of path length

Existing MPA algorithms only extract the most important path from the network. However, there can be many main streams in a field of study and ideas tend to merge or split over time. A single main path approach may not adequately convey the merge or split phenomena in a discipline's evolution. For representing these academic phenomena in our main path analysis, we adopted the longest path algorithm and multiple path extraction to extract papers that are more representative. We attempted to extract the main path that covered throughout the publication years in a citation network and the longest path that could reflect the diameter of the collected citation network from the origin to the latest.

Next, we will summarize the longest path algorithm that we utilized in our Java package, which Sedgewick and Wayne (2011) provided. This algorithm uses three steps: (1) topological sort, (2) edge relaxation, and (3) multiple path extraction. We then create multiple paths by merging distinct detected paths.

Topological Sort This method arranges each node of a graph into a linear form prior to edge relaxation. We used a depth first search (DFS) algorithm for topological sorting. DFS starts at one arbitrary node, which is selected as the root, and explores all branches of the graph. We used a reverse post-order for our DFS ordering. Fig. 2 presents the arrangement of nodes in reverse post-order using the simple network Fig. 2(a). N1 from Fig. 2(a) was selected as the root and topological search started from the root and expanded to all branches. If a node with no branches or children was found, that node was added to the search in reverse order. Exploration then returned to the parent node to visit other branches or children. Once this process visits all nodes in a graph, the topological ordering is complete.

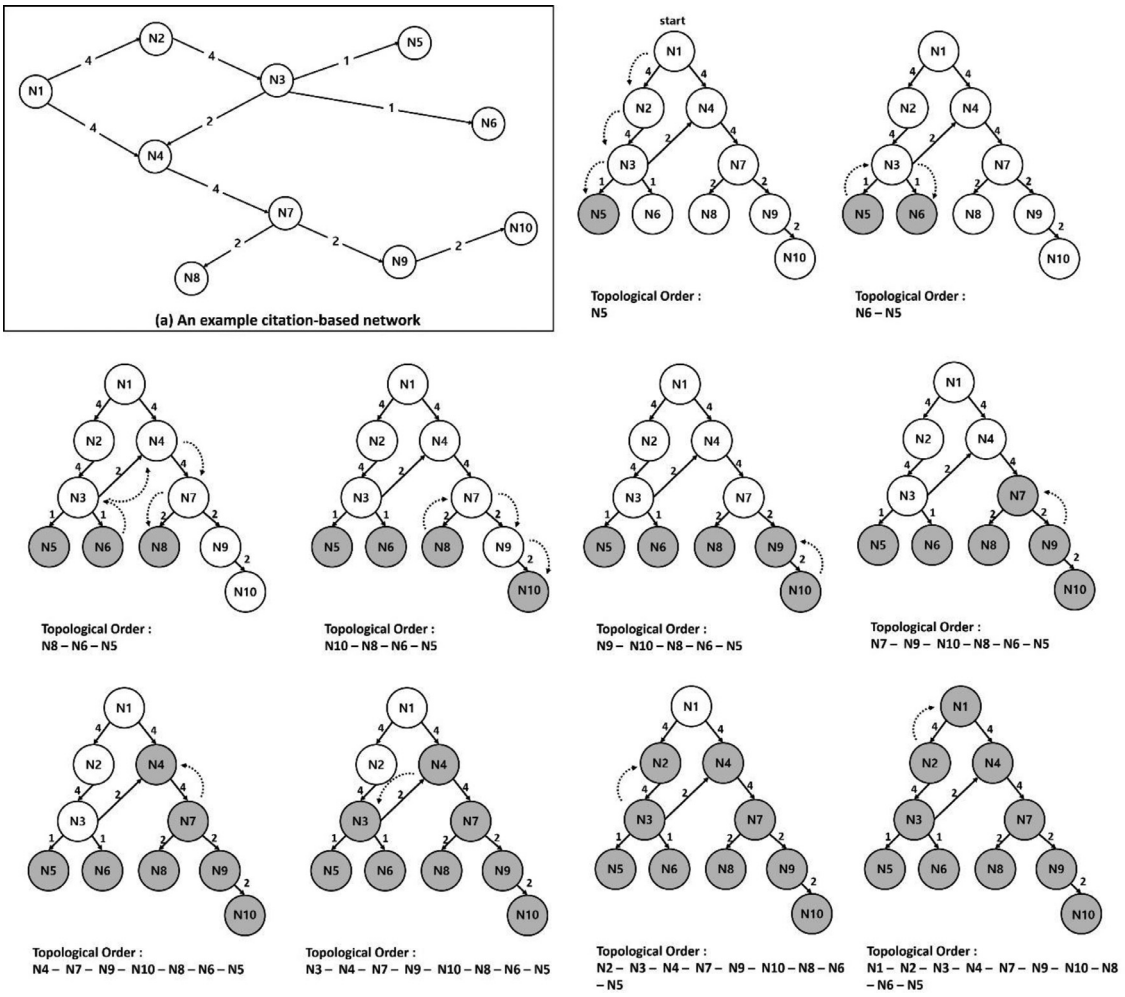


Fig. 2. The arrangement of nodes in reverse post-order using the simple network.

Edge relaxation After arranging all the nodes in the graph, for each edge, we determined whether the edge was valid or invalid as based on certain criteria. When distTo[w] was the sum of weights from the root node to node w, distTo[v] was the sum of weights from the root node to node v, and e.weight() was the edge weight between nodes v and w, the threshold for the validity of the edge between nodes v and w was defined as follows:

$$\text{distTo}[w] < \text{distTo}[v] + \text{e.weight}()$$

If the distTo[w] value was larger than the sum value of distTo[v] and e.weight(), the edge between nodes v and w was discarded because there was a path more weighted than the path containing the edge between nodes v and w.

Multiple path extraction After edge relaxation, only distinct paths from the sources to the targets are left. For Fig. 2(a), path N1-N2-N3-N4-N7-N9-N10 is extracted as the longest path. Therefore, we collected all sources and targets and found the paths from identified sources to targets. The obtained path lengths varied and we ranked the paths in order of length. We did not choose a single longest path, but created multiple paths by merging detected paths in order of path length.

Assuming there is another source (N11) in Fig. 2, as shown in Appendices, each node, N1 and N11, is set as a start node in edge relaxation before extraction of the longest path. Therefore, the following two different longest paths are generated: N1-N2-N3-N4-N7-N9-N10 and N11-N4-N7-N9-N10. We provided more detailed figures in Appendices that explain the first two of the three steps used in the example of a citation network containing two sources: topological sort and edge relaxation. Appendix A shows topological order. As explained in our study, we used a depth-first search (DFS) algorithm for topological sorting. DFS starts at one arbitrary node, which is selected as the root, and explores all branches of the graph. Even though there is one more source, a topological sort is performed once for all nodes. Appendix B shows edge relaxation when the citation network contains many sources after topological order. Each source, i.e., N1 and N11, is set as a start node in edge relaxation before extraction of the longest path. Please note that

the topological order is the same in the cases of N1 and N11; however, the process of edge relaxation is different between the start nodes N1 and N11.

3.5. Topic-integrating to the extracted path

Scientific knowledge transfers from a cited paper to a citing paper. In a citation chain, a piece of knowledge in the preceding paper is present in some way in the succeeding paper (Verspagen, 2007). Knowledge disseminates to multiple citing papers, converges in a single citing paper, or shifts from cited papers to other topics. However, the traditional MPA does not consider topic or subject flows, thus researchers must examine each paper in the extracted main path to understand its topic dynamics. Therefore, we attempt to illuminate topics through empirical analysis so that we can trace topic dynamics automatically in the scientific trajectory.

Topic integration provides a more comprehensive exploration of the development trajectory in a field of study. This work carries out separately with extracting the path and we matched the papers of extracted paths with discovered topics of individual documents from the whole dataset we collected. For topic integration, we first uncovered the hidden topics in the dataset. We used LDA, which Blei et al. (2003) proposed. LDA is a statistical model that learns co-occurrence patterns and model topics based on the words present in a document collection. LDA finds not only the topic distribution of the entire collection but also that of individual articles (DiMaggio et al., 2013). For discovering topic distributions, the number of topics is set in advance. Iterative tests for experimental settings are required to find the appropriate number of topics. We set the number of topics based on the range of an exemplar discipline because the extent of a discipline's subfields determines the number of topics. After we determined the number of topics and obtained the topic distribution of both the entire collection and individual articles, we integrated the results of the topic models into the extracted multiple main paths. In particular, the topic-document matrix that LDA derived was used to select which topic was the most dominant in an article. In other words, each article was labeled as the topic that was assigned to the highest topical probability. As a result, each article node in our main path can be distinguished by its topic.

3.6. Analyzing the scientific trajectories

This section describes how we calculate three features of the scientific trajectories from the extracted paths: academic emergence, authority, and topic dynamics.

Academic emergence. After integrating topic data into the extracted longest paths, we find the papers with the earliest publication dates among the sources and search for the nodes where those earlier works converge. We call earliest papers cornerstone works and the consolidated works academic emergence.

Authority. Kleinberg (1999) suggested a way of refining the importance of pages on the Web, which referred to a certain type of page as an authoritative page. His-formulation adopted eigenvectors. Authorities are defined as articles cited by many other articles, while hubs are defined as articles citing many other articles. Batagelj (2003) added that quality authorities are articles that are cited by quality hubs, and vice versa. In other words, these nodes received many citations and cited many other highly cited papers. Knowledge convergence occurs because the paper cites many authoritative papers and knowledge divergence occurs because the paper is cited by many other papers. We measure the authoritative score of each work in our entire citation network, a process implemented in Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

Topic dynamics. After integrating topic data into the extracted longest paths, we reconfigure the longest paths according to chronicle order to follow topic flows and interaction among topics. We analyzed topic dynamics at path level among extracted longest path clusters.

4. Results

4.1. Traditional key-route main path

For the traditional main path, we extracted the key-route main path by applying the SPC weight implemented in Pajek software. The key-route method chooses the top rank of the largest weights overall and provides multiple main paths by choosing the top N ranks (Liu & Lu, 2012). We extracted the top five key-routes by applying local and global approaches in analyzing the traditional main path.

Fig. 3 shows the key-route main paths on the top five SPC ranks. We marked the five largest SPC routes with red lines in Figs. 3(a) and 3(b). The local key-routes presented 267 main paths including 85 nodes, as shown in Fig. 3(a), while the global key-routes presented three main paths including 44 nodes, as shown in Fig. 3(b). We found that the local key-routes and the global key-routes shared 30 identical nodes and 27 identical links between them.

The local key-routes showed six sources published from 1966 to 1973 and 27 sinks published from 2014 to 2016; however, the global key-routes showed one source published in 1993 and one sink in 2016. Therefore, the six sources in the local key-routes represent instances of academic emergence in healthcare informatics (Fig. 3(a)). These six sources consolidated in Harris' work (PubMed ID 1,225,865) titled "Effect of population and health care environment on hospital utilization" published in the journal Health Services Research, 10(3), in 1975. (Health Services Research, 10(3), 229–243) published in 1975. From this result, we can infer the field of healthcare informatics began with Harris' work in 1975 (academic emergence), which was affected by six papers (cornerstones).

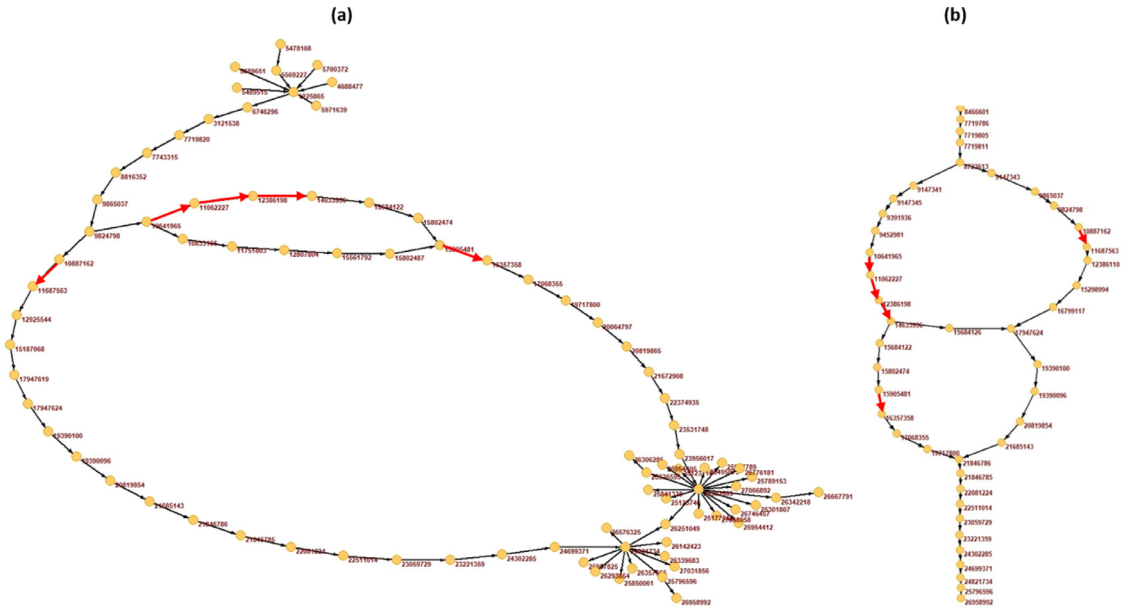


Fig. 3. The main paths produced by the key-route algorithm: (a) local key-routes, (b) global key-routes. The node labels indicates the PubMed ID (PMID) and the red line presents the five largest SPCs.

The local and global key-routes showed divergence, as well as convergence, at certain points along the key-routes. However, representing by less or more a hundred papers in the main path is not enough to capture the detailed knowledge transfer patterns of a mature discipline such as healthcare informatics, which is characterized by an immense number of publications and a wide span of history in the real world. Our collected dataset included a total of 274,297 individual papers and 595,548 citing–cited pairs published over a period of more than four decades. In our analysis of the trajectory of healthcare informatics, the longest path that we extracted shows the four most significant path clusters that contain over 200 nodes each, representing 0.29% of the entire dataset.

4.2. Topic-integrated path extraction

4.2.1. Extraction of the longest paths

Fig. 4 illustrates the path extracted by our degree-weighted longest path approach. The seed document set in our study includes documents published from 1970 to 2017. As shown in Fig. 4, our main paths' characteristic results are different from those of the SPC-weighted key-route main paths in Fig. 3 because the earliest publication year was 1966 and the latest publication year was 2017. The average number of nodes in four path clusters was 207 and the average number of sources was 112. The four largest SPC routes in Fig. 3 appeared in our extracted longest paths, as shown in Figs. 4(a) and 4(d); the top three SPC routes also appeared, shown in Fig. 4(d), as did the 4th largest SPC route, shown in Fig. 4(a); however, only one pair from the 5th largest SPC route was found in our extracted path, as shown in Fig. 4(b), and the other pair was not found in our longest paths. Therefore, we learned that our longest paths expanded the 5 largest SPC routes to three different path clusters and found one more path cluster that the traditional SPC-based MPA had not found.

4.2.2. Topics discovered in healthcare informatics

Healthcare informatics is a multidisciplinary field that incorporates information and computer technology into health and biomedicine to improve healthcare services (Mettle & Raptis, 2012; Sweeney, 2017). It is also called health informatics, medical informatics, nursing informatics, clinical informatics, or biomedical informatics (Nadri et al., 2017). For topic-integrated path analysis, we first discovered the topics in our healthcare informatics dataset using the LDA algorithm. To identify the most distinctive topics, the appropriate number of topics, k, should be specified in advance. To set a value k, LDA determines the results' quality. A small number of k causes the result in broader topics containing very common words, while the granular and overfitted topics were obtained as a large number of topics is set (Blair et al., 2020). Goodness-of-fit for the statistical measures for topic models, such as held-out likelihood, perplexity (Blei et al., 2003), semantic coherence (Mimno et al., 2011), and empirical knowledge by domain experts, is suggested to select an appropriate k in practice.

In our study, we went over healthcare informatics' intellectual structures that the related studies displayed, as shown in Table 1. Raghupathi and Nerur (2008; 2010) and Chen et al. (2014) suggested 13 to 14 topics, and Chen et al. (2019) suggested 22 topics to the intellectual structure of healthcare informatics. We took the number of topics in Table 1 into account. Meanwhile, we determined that

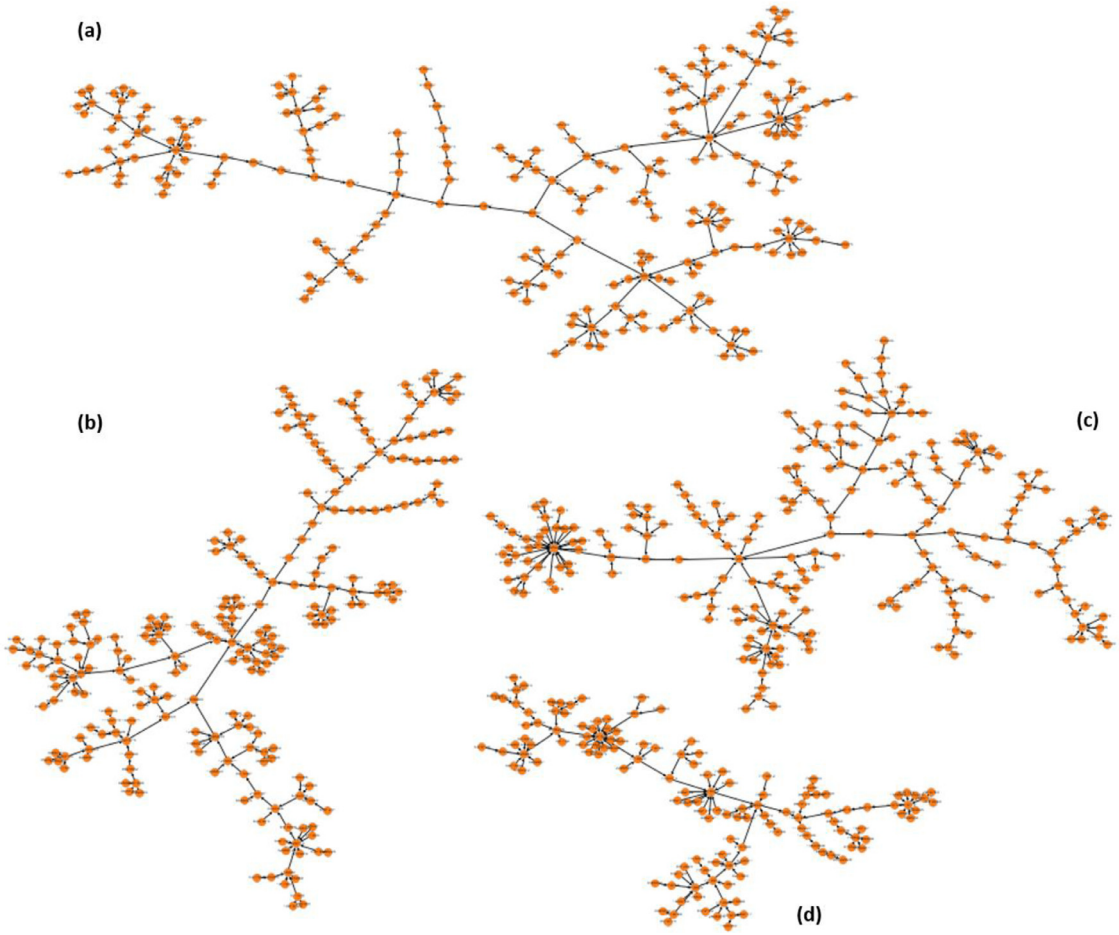


Fig. 4. The path clusters by the longest path algorithm with degree centrality.

20 was the optimal k of topics after the iterative tests for the experimental setting in the range of 10 to 30. Biology and information science experts reviewed the LDA topic modeling results and labeled topic names in light of the extent of healthcare informatics subfields.

Table 2 shows the topic modeling results. The topics uncovered in healthcare informatics included those such as "health survey (T16)" and "clinical practice (T18)," which accounted for the largest proportions (42.1% and 34.3%, respectively) of the healthcare informatics field, as shown in Fig. 5. Furthermore, "smoking cessation (T7)" and "palliative care (T11)" were also weighted subjects (both 8.8%) from the topic modeling results. The remaining topics each accounted for less than 1% of the field.

We integrated the results of the 20 topics into the extracted longest paths called topic-integrated main paths. Fig. 5 shows the results of LDA on two sides in our study. The left pie chart in Fig. 5 shows topic proportions over the entire dataset as described above, while the right pie chart shows topic proportions over publications belonging to the main paths we extracted.

We found that three major topics appeared in the topic-integrated main paths. Unlike the composition of topics in the entire dataset, T0 (clinical decision support system) was the most popular topic in the main paths (29.5% of nodes). This topic was positioned at a low rank (14/20) among the topics of the entire dataset. The second-most prominent topic was T18 (clinical practice, 22.4%) and the third was T15 (primary health care, 10.8%), followed by T10 (stroke, 7.4%), T16 (health survey, 6.6%), and T8 (medical care, 6.6%). From this ranking, we learned that the majority of research in healthcare informatics focuses on applied studies, including clinical system development, application, and implementation.

5. Exploring scientific trajectories

We redrew the main paths' results in Fig. 4 by projecting two panels (topic and time) in Fig. 6, such that nodes and topics are presented in chronological order. The x-axis represents the 20 topics and the y-axis represents the publication year. Figs. 6(a) to 6(d) correspond to four clusters in Fig. 4, respectively. We redrew Fig. 4 because it was easy to find the earliest papers and the interaction

Table 1
Topics in healthcare informatics-related fields.

No.	Raghupathi and Nerur (2008)	Raghupathi and Nerur (2010)	Chen et al. (2014)	Chen et al. (2019)
1	HIS Evaluation	Communication and eHealth	Security of HIT	Health IS Implementation 80
2	Communication and eHealth	Medical Imaging Technology	Implications of HIT	Health IS Acceptance
3	E-Health	A.I./Decision Support	Medical Information Retrieval	Health IS-Induced Anxiety and Resistance
4	Clinical DSS	Ontology and medical terminology	Medical Image Processing and Management	Health IS Productivity
5	Adoptions, Outcome, and Policy	Remote Monitoring, Mobile Computing	Trust in HIT	Health IS Outsourcing, Performance, and Investment
6	Telemedicine and Communication	User Acceptance Quality	EMR and HER	Health IS Innovation
7	Mobile Computing	Bioinformatics	Knowledge Management in Healthcare	National Health IS Programs
8	Internet- and Web-based Health Care	Clinical Information Systems	TAM of HIT	Security of Health IS
9	Quality and Integration	User Interface	National HIT Programs	Health Information Interchange
10	Public Health Informatics	Natural Language Processing	General HIT Applications	Health IS Compliance
11	Use and Impact of HIS	Health Informatics	HIT Innovation	Trust of Health IS
12	Medical Safety	e-Health	HIT and Organizations	Health IS and Patient-Centered Care
13	Health Policy, Quality of Care	Computational Genomics	Clinical Decision Support	EMR and EHR
14		Analysis and extraction in genomics	Telemedicine	Mobile Health
15				Health Analytics and Data Mining
16				Health Information Search and Retrieval
17				Health Image Retrieval and Management
18				Clinical Pathway and Treatment Management
19				Knowledge Management in Healthcare
20				RFID and Tracking in Healthcare
21				Health Consumer Privacy
22				Online Health Communities and Digital Services

among topics in individual clusters. In Fig. 6, the nodes represent papers and the links denote the citation relationships between neighboring nodes. Our links are directional arcs where the arrow points from the cited paper to the citing paper, meaning scientific knowledge disseminates along the link direction.

In the healthcare informatics field, we found four main path clusters (hereafter, mainstream or MS) as shown in Figs. 4 and 6. Each MS of the longest paths discovered contained interesting characteristics, and we attempted to characterize those four MSs based on the following three types of knowledge diffusion: (1) emergence, (2) authority, and (3) topic dynamics. Detailed descriptions of these characteristics are provided below. We call Fig. 6(a)–(d) 1MS to 4MS, respectively.

5.1. Academic emergence

From the perspective of development trajectories, healthcare informatics began developing with the advent of computer technologies in the 1970s (Cesnik & Kidd, 2010). Therefore, a paper in the citation network that was published in the 1960s may be a cornerstone from which healthcare informatics emerges. Academic emergence occurs at the consolidation point where the cornerstones converge.

Table 3 reports that the topic-integrated longest main path found twelve cornerstones in the 1960s and academic emergence in the 1970s. Four of these cornerstones have the topic of medical care (T8). This implies that the field of healthcare informatics arose mainly from efforts related to medical care. Five articles were found at the consolidated points, and their associated topics were diagnostic test (T1), medical care (T8), stroke (T10), primary health care (T15), and clinical practice (T18). In detail, 10 cornerstone articles appeared in the 4MS (Fig. 6(d)), then merged into four emergent articles, whereas two cornerstone articles appeared in the 2MS (Fig. 6(b)), then were connected to two emergent articles through one article each. Therefore, we can conclude that the 4MS was more relevant to the emergence of healthcare informatics in the mid-1960s and the 1970s.

5.2. Authority

We measured the quality of nodes in the extracted longest paths based on the formulation of authority that Kleinberg (1999) proposed and we discovered eight authoritative articles in our main paths. These authoritative articles act as hubs in the network.

Table 2
Topic discovered in healthcare informatics.

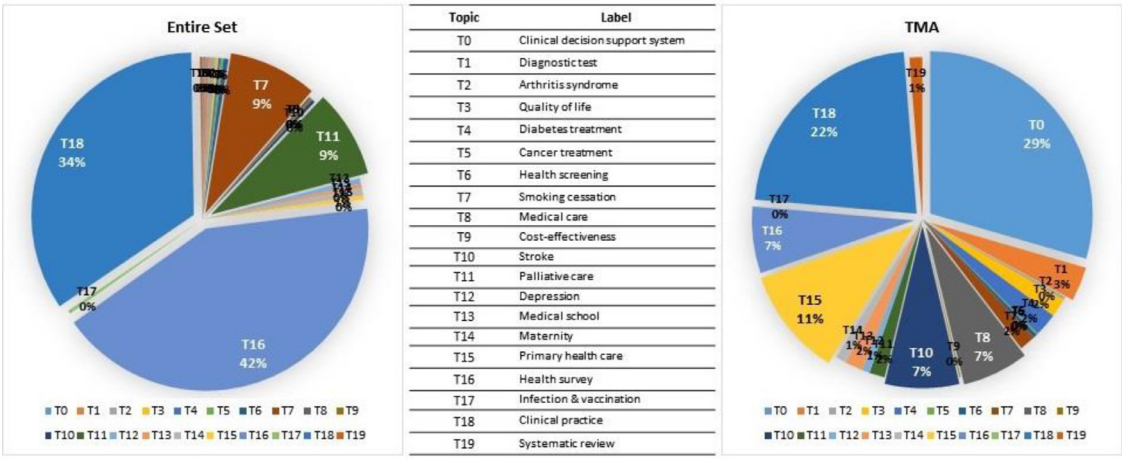
Topic	T0 clinical decision support system	T1 diagnostic test	T2 arthritis syndrome	T3 QOL (quality of life)	T4 diabetes treatment
Top Words	clinical	test	disease	scores	drug
	medical	regression	surgery	health	diabetes
	systems	estimates	cancer	quality	medication
	electronic	diagnostic	pain	validity	adherence
	patient	time	syndrome	scale	prescription
	records	statistical	clinical	reliability	therapy
	analysis	bias	knee	life	pharmacy
	support	sample	risk	measures	asthma
	classification	risk	therapy	assessment	hypertension
	decision	sensitivity	hip	instrument	blood
Topic	T5 cancer treatment	T6 health screening	T7 smoking cessation	T8 medical care	T9 cost-effectiveness
Top Words	cancer	cancer	trial	health	cost
	pain	women	intervention	care	cost-effectiveness
	chemotherapy	risk	randomized	insurance	economic
	oral	screening	program	Medicare	clinical
	therapy	breast	protocol	services	therapy
	symptom	men	clinical	costs	health
	breast	age	effectiveness	Medicaid	QALY
	lung	factors	care	cost	life
	nausea	years	smoking	coverage	disease
	opioid	disease	cessation	utilization	benefits
Topic	T10 stroke	T11 palliative care	T12 depression	T13 medical education	T14 maternity
Top Words	patient	patient	life	medical	health
	mortality	care	quality	students	children
	risk	palliative	depression	medicine	care
	heart	patient	physical	education	women
	acute	cancer	symptoms	training	countries
	surgery	decision	mental	clinical	HIV
	discharge	family	adults	learning	community
	failure	physicians	chronic	teaching	parents
	admission	qualitative	social	skills	maternal
	coronary	communication	anxiety	school	birth
Topic	T15 primary health care	T16 health survey	T17 infection & vaccination	T18 clinical practice	T19 systematic review
Top Words	care	survey	HIV	health	review
	patient	participants	infection	care	systematic
	primary	online	chronic	development	evidence
	physician	internet	influenza	implementation	trials
	quality	response	vaccination	clinical	research
	medical	users	pulmonary	practice	clinical
	nursing	respondents	respiratory	policy	literature
	home	questions	hepatitis	process	quality
	visits	literacy	COPD	medical	interventions
	satisfaction	questionnaire	antibiotic	public	reporting

The eight authority points that we discovered are marked in red in each MS in Fig. 6. The knowledge convergences and divergences occurred vigorously during one decade between the mid-1990s and mid-2000s as shown in Fig. 6. The eight authorities are listed in Table 4. Two authorities were observed in the 1MS, four authorities in the 2MS, and two authorities in the 4MS. However, there was no authority observed in the 3MS.

5.3. Topic dynamics

Various development trajectories in healthcare informatics can be observed from the four main flows of knowledge in the field.

The 1MS begins with T15 in 1971 and ends with T4 in 2017. The topics vary between five choices: T0, T8, T10, T15, and T18. T0 (Clinical decision support system) is the most dominant topic and remained as the most highly weighted topic from the 1990s



Topic	Label
T0	Clinical decision support system
T1	Diagnostic test
T2	Arthritis syndrome
T3	Quality of life
T4	Diabetes treatment
T5	Cancer treatment
T6	Health screening
T7	Smoking cessation
T8	Medical care
T9	Cost-effectiveness
T10	Stroke
T11	Palliative care
T12	Depression
T13	Medical school
T14	Maternity
T15	Primary health care
T16	Health survey
T17	Infection & vaccination
T18	Clinical practice
T19	Systematic review

Fig. 5. Comparison of the 20 topics' proportions between the entire collection and the topic-integrated longest paths.

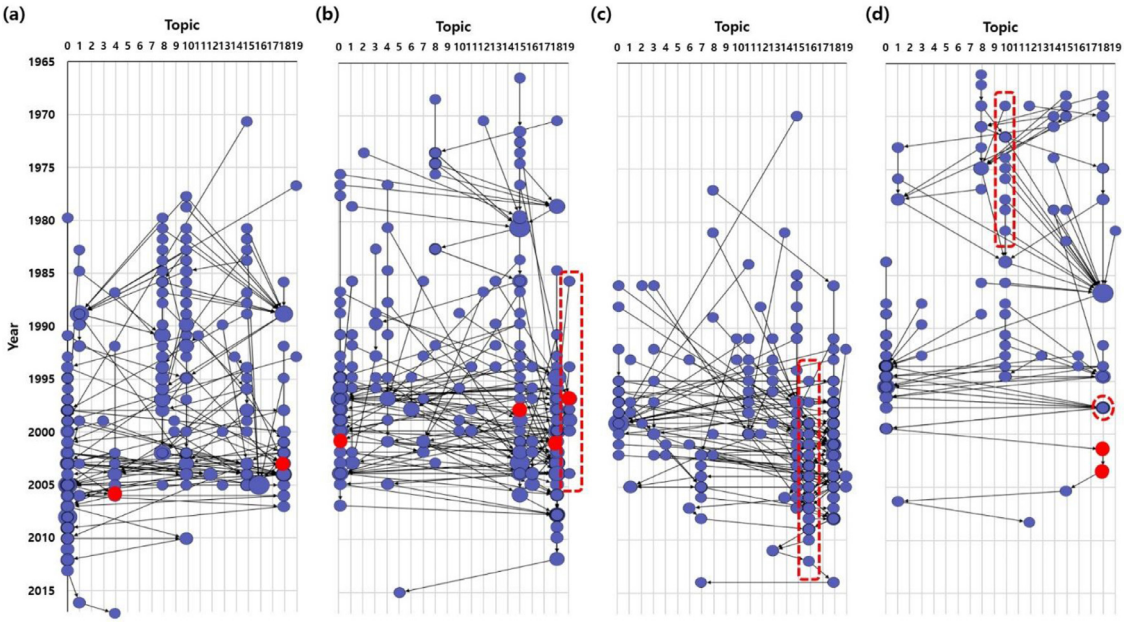


Fig. 6. Knowledge flows in topic-integrated longest path of healthcare informatics. Reconfiguration of the data in Fig. 4 by topic (x-axis) and time (y-axis).

to the 2010s. In the 1MS, convergence occurred in 2005. The article titled "Effect of CPOE user interface design on user-initiated access to educational and patient information during clinical care (Rosenbloom et al. 2005)," whose topic was T16 (health survey) (colored as a red node in Fig. 6(a)) was influenced by six topics (T0, T9, T10, T13, T15, and T18). Furthermore, this convergent knowledge shifted in the same year when Miller et al. (2005) cited this paper. T16 for Rosenbloom et al. (2005) shifted to T0 for Miller et al. (2005).

In the 2MS, there is knowledge exchange among three main topics (T0, T15, and T18) from 1967 to 2015. The 2MS (Fig. 6(b)) shows the same three predominant topics (T0 [23.2%], T15 [15.4] and T18 [31.6%]). These three topics account for 70% of the 2MS. This stream originated in 1967 with an article titled "A structural model for the patient care operation," which was authored by Jelinek. Its topic is primary health care (T15). At the beginning of the MS, T15 evolved and developed based on the influence from T1, T8, T12, and T18 until the 1970s. Unlike T15, T0 evolved in 1976 when the article titled "AUTOGRP: an interactive computer

Table 3
Twelve cornerstones and the emergence of healthcare informatics. 1MS to 4MS indicate Figs. 6(a) to 6(d) respectively.

Cornerstone PMID (Topic No.) and Publication Year	Consolidation	Academic Emergence Authors and Title	MS
6,081,241 (T15), 1967	(via 5,072,860 (T15)) →6,782,043 (T15)	Cavaiola, L. J., & Young, J. P. (1980). An integrated system for patient assessment and classification and nurse staff allocation for long term care facilities. *Health Services Research*, 15(3), 281.	2MS
5,377,856 (T8), 1969	(via 1,205,865 (T8)) →116,990 (T18)	Vraciu, R. A. (1979). Programming, budgeting, and control in health care organization: the state of the art. *Health Services Research*, 14(2), 126.	
5,971,639 (T8), 1966 5,700,372 (T15), 1968 5,889,651 (T10), 1969 4,979,042 (T15), 1969	→1,225,865 (T8)	Harris, D. M. (1975). Effect of population and health care environment on hospital utilization. *Health Services Research*, 10(3), 229.	
6,054,380 (T8), 1967	(via 5,377,854 (T8)) →4,631,546 (T10)	Edwards, M., Miller, J. D., & Schumacher, R. (1972). Classification of community hospitals by scope of service: Four indexes. *Health Services Research*, 7(4), 301.	4MS
5,700,373 (T18), 1968	(via 5,095,658 (T8)) →4,631,546 (T10)		
5,377,854 (T8), 1969	→4,631,546 (T10)		
5,809,795 (T18), 1969	→632,103 (T1)	Brooks, C. H. (1978). Infant mortality in SMSAs before Medicaid: test of a causal model. *Health Services Research*, 13(1), 3.	
5,809,796 (T12), 1969 5,889,653 (T18), 1969	→4,986,435 (T18)	Berg, R. L., Browning, F. E., Hill, J. G., & Wenkert, W. (1970). Assessing the health care needs of the aged. *Health Services Research*, 5(1), 36.	

Table 4
Authoritative papers that act as hubs in topic-integrated longest path. 1MS to 4MS indicate Fig. 6(a)–(d), respectively.

Authors and Title	Topic No.	MS
Harpole, L. H., Khorasani, R., Fiskio, J., Kuperman, G. J., & Bates, D. W. (1997). Automated evidence-based critiquing of orders for abdominal radiographs: Impact on utilization and appropriateness. *Journal of the American Medical Informatics Association*, 4(6), 511–521.	T19	2MS
Shojania, K. G., Yokoe, D., Platt, R., Fiskio, J., Ma'Luf, N., & Bates, D. W. (1998). Reducing vancomycin use utilizing a computer guideline: Results of a randomized controlled trial. *Journal of the American Medical Informatics Association*, 5(6), 554–562.	T15	2MS
Murff, H. J., & Kannry, J. (2001). Physician satisfaction with two order entry systems. *Journal of the American Medical Informatics Association*, 8(5), 499–509.	T0	2MS
Patterson, E. S., Cook, R. I., & Render, M. L. (2002). Improving patient safety by identifying side effects from introducing bar coding in medication administration. *Journal of the American Medical Informatics Association*, 9(5), 540–553.	T18	2MS
Effken, J. A., & Carty, B. (2002). The era of patient safety: Implications for nursing informatics curricula. *Journal of the American Medical Informatics Association*, 9(Supplement_6), S120–S123.	T18	4MS
Bates, D. W., Kuperman, G. J., Wang, S., Gandhi, T., Kittler, A., Volk, L., … & Middleton, B. (2003). Ten commandments for effective clinical decision support: Making the practice of evidence-based medicine a reality. *Journal of the American Medical Informatics Association*, 10(6), 523–530.	T18	1MS
Ash, J. S., Berg, M., & Coiera, E. (2004). Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. *Journal of the American Medical Informatics Association*, 11(2), 104–112.	T18	4MS
Van Der Sijs, H., Aarts, J., Vulto, A., & Berg, M. (2006). Overriding of drug safety alerts in computerized physician order entry. *Journal of the American Medical Informatics Association*, 13(2), 138–147.	T4	1MS

system for the analysis of health care data" by Mills et al. (1976) was published and stood alone until the mid-1990s. T18 originated in 1971, but merged into T15 in 1980 via T8. Therefore, T18 actually formed after 1985 when the article titled "Physician utilization: The state of research about physicians' practice patterns" by Eisenberg was published. Since the mid-1980s, intellectual interchange occurred actively between T0, T15, T18, and others. However, only T8 remained as the most highly weighted topic in the 2010s in the 2MS. Studies on T19 were only highlighted in the 2MS in healthcare informatics (dotted line in Fig. 6(b)).

In the 3MS, we found four main topics (T0, T15, T16, and T18) and other minor topics from 1970 to 2014. The 3MS (Fig. 6(c)) primarily focuses on T15, T16, and T18. At the beginning of the MS, knowledge regarding T15 (Penchansky & Fox, 1970), T8 (Berki et al., 1977; McGuire, 1981), and T14 (Pauker et al., 1981) transferred to T0 (Westberg & Miller, 1999), T18 (Luft 1986), T16 (Booske et al., 1999), and T11 (Brennan & Strombom, 1998), respectively. Therefore, it is appropriate to say that T15 developed since the article titled "Measurement of physician performance by standardized patients: Refining techniques for undetected entry in physicians' offices" that (Woodward et al., 1985) published in 1985. T16 shows notable development in the 3MS compared to the other main streams (shown as a dotted line in Fig. 6(c)). T16 accounts for 6.6% of the extracted multiple main paths, but is mostly found in the 3MS.

In the 4MS, knowledge chains occurred mainly in T0, T10, and T18, with the earliest paper detected in 1966. T10 has actively evolved since the 1980s, as shown in the 1MS (Fig. 6(a)). In contrast, we can see that the dynamics of T10 were captured during the mid-1960s and 1970s in the 4MS (shown as a dotted line in Fig. 6(d)). A similar pattern can be seen for T8, where vigorous research occurred in the 1980s to 2000s in the 1MS, but appeared earlier in the mid-1960s and 1970s in the 4MS. Additionally, a ping-ponged interaction exists between T0 and T18 in the 4MS. For example, the recurring knowledge flows followed cited-citing relationships as follows: T0 (five papers) → T18 (Cimino, 1998; Chute et al., 1998 in a red dotted circle) → T0 (Bakken et al., 2000) → T0 (Harris et al., 2000) → T18 (Effken & Carty, 2002).

6. Conclusion

Tracing knowledge diffusion in a research field plays a pivotal role in understanding the dynamics of the field and its impacts on related scientific communities. To adapt the benefits of MPA approaches to multidisciplinary domains, we proposed topic-integrating path analysis by considering three facets of trajectory analysis: emergence, authority, and topic dynamics.

This study's contributions are twofold. First, we claim that a small-scale dataset results in main paths with relatively short lengths of paths. We exploit the longest path algorithm, which searches paths exhaustively, enabling the identification of more sequences of knowledge transfer in the large-scale datasets. Real-world knowledge flows are revealed in our new topic-integrated longest paths, allowing us to understand the phenomena of an academic emergence in an interdisciplinary or multidisciplinary discipline, meaning the longest paths that we detected more fully capture the detailed knowledge transfer patterns of the mature, multidisciplinary fields such as healthcare informatics, which is characterized by an immense number of publications and a wide span of history in the real world. Second, this study revealed the knowledge development tracks of topics in the healthcare informatics field. Topic-integrated longest paths highlight important topics that move from the main path to extended subfields.

One limitation of topic-integrated path analysis is that algorithms for weighted paths are typically based on social network measures, which may put more weight on direct citations in a citation network. For future research, we will extensively explore optimal measures for weighted paths in citation networks. We also plan to apply topic-integrated path analysis to other domains, such as information science and electrical engineering, to examine how the results of topic-integrated path analysis in different domains are similar or different to those reported in this study. This cross-domain analysis may help determine whether a specific domain's characteristics influence knowledge diffusion in a given field.

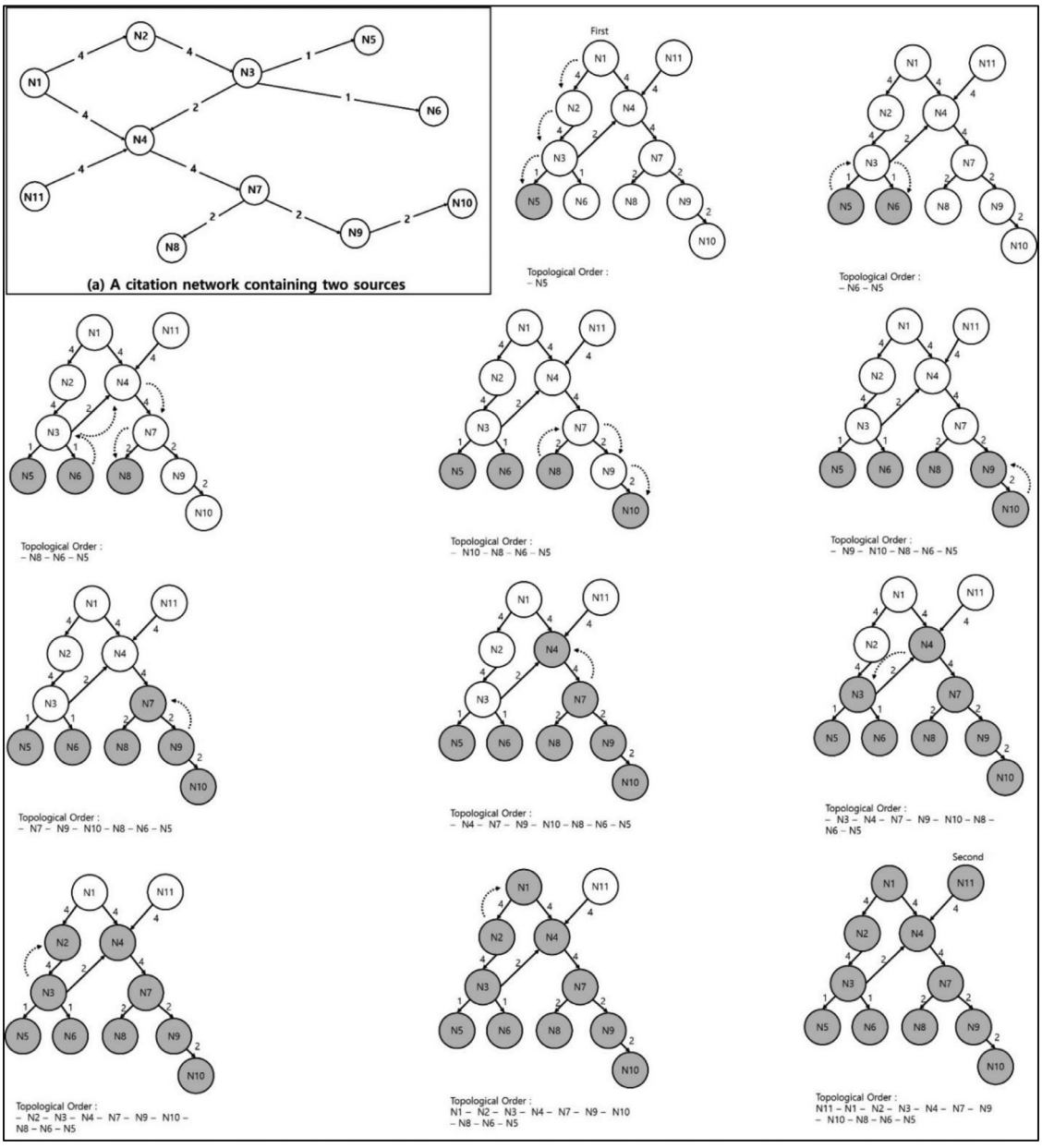
CRediT authorship contribution statement

Erin H.J. Kim: Conceptualization, Formal analysis, Data curation, Writing – original draft. **Yoo Kyung Jeong:** Conceptualization, Formal analysis, Data curation, Writing – original draft. **YongHwan Kim:** Conceptualization, Formal analysis, Data curation, Writing – original draft. **Min Song:** Conceptualization, Formal analysis, Data curation, Writing – original draft.

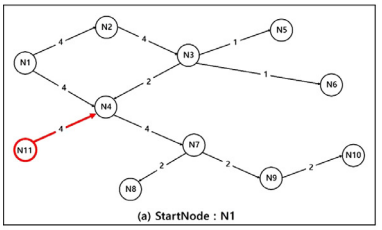
Acknowledgement

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2019S1A5A8033713). This work was partly supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2020–0–01361, Artificial Intelligence Graduate School Program (Yonsei University)).

Appendix A. The step of Edge relaxation in the simple citation network containing two sources.


(a) A citation network containing two sources
First
Second
Topological Order : – N5
Topological Order : – N6 – N5
Topological Order : – N8 – N6 – N5
Topological Order : – N10 – N8 – N6 – N5
Topological Order : – N9 – N10 – N8 – N6 – N5
Topological Order : – N7 – N9 – N10 – N8 – N6 – N5
Topological Order : – N4 – N7 – N9 – N10 – N8 – N6 – N5
Topological Order : – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5
Topological Order : – N2 – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5
Topological Order : N1 – N2 – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5
Topological Order : N11 – N1 – N2 – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5

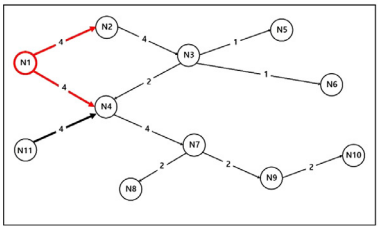
Appendix B. Topics in healthcare informatics-related fields.



(a) StartNode : N1

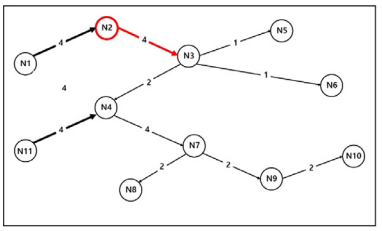
Topological Order :
N11 – N1 – N2 – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5

Node	edgeTo	new_edgeTo	distTo[w]	new_distTo[w]
N1			0	
N2				
N3				
N4	N11->N4		4	
N5				
N6				
N7				
N8				
N9				
N10				
N11			0	



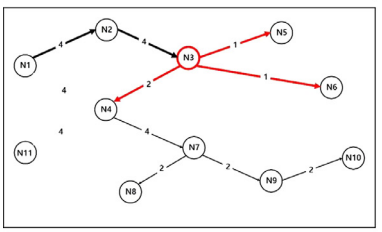
Topological Order :
N11 – N1 – N2 – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5

Node	edgeTo	new_edgeTo	distTo[w]	new_distTo[w]
N1			0	
N2	N1->N2		4	
N3				
N4	N11->N4	N1->N4	4	4
N5				
N6				
N7				
N8				
N9				
N10				
N11			0	



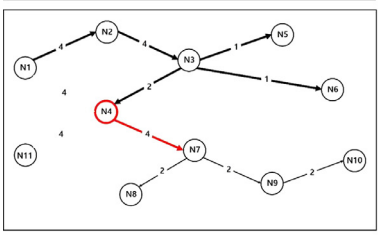
Topological Order :
N11 – N1 – N2 – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5

Node	edgeTo	new_edgeTo	distTo[w]	new_distTo[w]
N1			0	
N2	N1->N2		4	
N3	N2->N3		4	
N4	N11->N4		4	
N5				
N6				
N7				
N8				
N9				
N10				
N11			0	



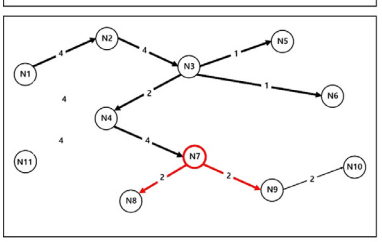
Topological Order :
N11 – N1 – N2 – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5

Node	edgeTo	new_edgeTo	distTo[w]	new_distTo[w]
N1			0	
N2	N1->N2		4	
N3	N2->N3		4	
N4	N11->N4	N3->N4	4	10
N5	N3->N5		9	
N6	N3->N6		9	
N7				
N8				
N9				
N10				
N11			0	



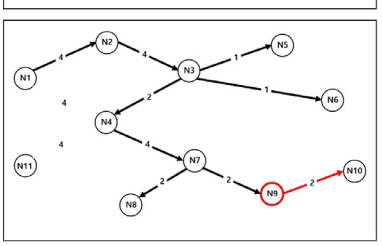
Topological Order :
N11 – N1 – N2 – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5

Node	edgeTo	new_edgeTo	distTo[w]	new_distTo[w]
N1			0	
N2	N1->N2		4	
N3	N2->N3		4	
N4	N3->N4		10	
N5	N3->N5		9	
N6	N3->N6		9	
N7	N4->N7		14	
N8				
N9				
N10				
N11			0	



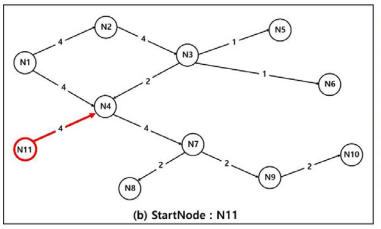
Topological Order :
N11 – N1 – N2 – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5

Node	edgeTo	new_edgeTo	distTo[w]	new_distTo[w]
N1			0	
N2	N1->N2		4	
N3	N2->N3		4	
N4	N3->N4		10	
N5	N3->N5		9	
N6	N3->N6		9	
N7	N4->N7		14	
N8	N7->N8		16	
N9	N7->N9		16	
N10				
N11			0	



Topological Order :
N11 – N1 – N2 – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5

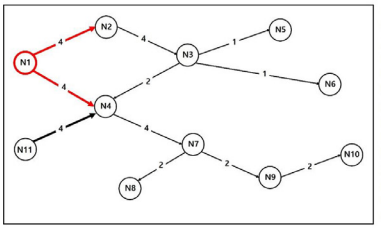
Node	edgeTo	new_edgeTo	distTo[w]	new_distTo[w]
N1			0	
N2	N1->N2		4	
N3	N2->N3		4	
N4	N3->N4		10	
N5	N3->N5		9	
N6	N3->N6		9	
N7	N4->N7		14	
N8	N7->N8		16	
N9	N7->N9		16	
N10	N9->N10		18	
N11			0	



(b) StartNode : N11

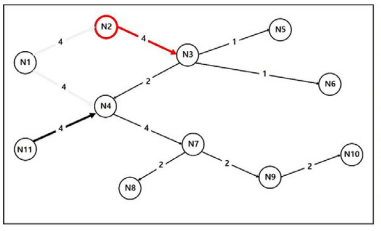
Topological Order :
N11 – N1 – N2 – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5

Node	edgeTo	new_edgeTo	distTo[w]	new_distTo[w]
N1			0	
N2				
N3				
N4	N11->N4		4	
N5				
N6				
N7				
N8				
N9				
N10				
N11			0	



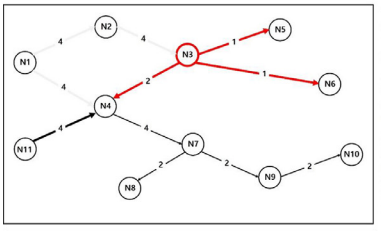
Topological Order :
N11 – N1 – N2 – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5

Node	edgeTo	new_edgeTo	distTo[w]	new_distTo[w]
N1			0	
N2	N1->N2		0	
N3				
N4	N11->N4	N1->N4	4	4
N5				
N6				
N7				
N8				
N9				
N10				
N11			0	



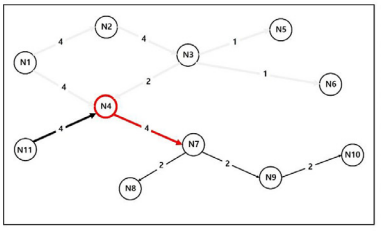
Topological Order :
N11 – N1 – N2 – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5

Node	edgeTo	new_edgeTo	distTo[w]	new_distTo[w]
N1			0	
N2				
N3	N2->N3		0	
N4	N11->N4		4	
N5				
N6				
N7				
N8				
N9				
N10				
N11			0	



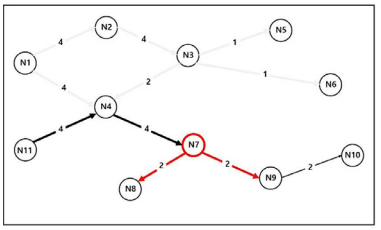
Topological Order :
N11 – N1 – N2 – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5

Node	edgeTo	new_edgeTo	distTo[w]	new_distTo[w]
N1			0	
N2				
N3				
N4	N11->N4	N3->N4	4	0
N5	N3->N5		0	
N6	N3->N6		0	
N7				
N8				
N9				
N10				
N11			0	



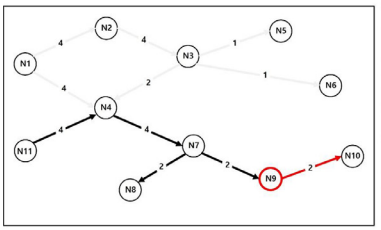
Topological Order :
N11 – N1 – N2 – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5

Node	edgeTo	new_edgeTo	distTo[w]	new_distTo[w]
N1			0	
N2				
N3				
N4	N11->N4		4	
N5				
N6				
N7	N4->N7		8	
N8				
N9				
N10				
N11			0	



Topological Order :
N11 – N1 – N2 – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5

Node	edgeTo	new_edgeTo	distTo[w]	new_distTo[w]
N1			0	
N2				
N3				
N4	N11->N4		4	
N5				
N6				
N7	N4->N7		8	
N8	N7->N8		10	
N9	N7->N9		10	
N10				
N11			0	



Topological Order :
N11 – N1 – N2 – N3 – N4 – N7 – N9 – N10 – N8 – N6 – N5

Node	edgeTo	new_edgeTo	distTo[w]	new_distTo[w]
N1			0	
N2				
N3				
N4	N11->N4		4	
N5				
N6				
N7	N4->N7		8	
N8	N7->N8		10	
N9	N7->N9		10	
N10	N9->N10		12	
N11			0	

Continued

References

Bakken, S., Cashen, M. S., Mendonca, E. A., O'Brien, A., & Zieniewicz, J. (2000). Representing nursing activities within a concept-oriented terminological system: Evaluation of a type definition. *Journal of the American Medical Informatics Association, 7*(1), 81–90.

Batagelj, V. (2003). Efficient algorithms for citation network analysis. arXiv preprint cs/0309023

Berki, S. E., Ashcraft, M., Penchansky, R., & Fortus, R. S. (1977). Enrollment choice in a multi-HMO setting: The roles of health risk, financial vulnerability, and access to care. *Medical Care*, 95–114.

Blair, S. J., Bi, Y., & Mulvenna, M. D. (2020). Aggregated topic models for increasing social media topic coherence. *Applied Intelligence, 50*(1), 138–156.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Booske, B. C., Sainfort, F., & Hundt, A. S. (1999). Eliciting consumer preferences for health plans. *Health Services Research, 34*(4), 839.

Brennan, P. F., & Strombom, I. (1998). Improving health care by understanding patient preferences: The role of computer technology. *Journal of the American Medical Informatics Association, 5*(3), 257–262.

Brughmans, T. (2013). Networks of networks: a citation network analysis of the adoption, use, and adaptation of formal network techniques in archaeology. *Literary and Linguistic Computing, 28*(4), 538–562.

Calero-Medina, C., & Noyons, E. C. (2008). Combining mapping and citation network analysis for a better understanding of the scientific development: The case of the absorptive capacity field. *Journal of Informetrics, 2*(4), 272–279.

Cesnik, B., & Kidd, M. R. (2010). In *History of health informatics: A global perspective: 151* (pp. 3–8). Studies in health technology and informatics.

Chen, L. C., Shih, I. C., & Liu, J. S. (2020). Identifying the main paths of knowledge diffusion in the voice over internet protocol. *Journal of Internet Technology, 21*(1), 85–98.

Chen, L., Baird, A., & Straub, D. W. (2019). An analysis of the evolving intellectual structure of health information systems research in the information systems discipline. *Association for Information Systems, 20*(8), 1023–1074.

Chen, L., Baird, A., & Straub, D. (2014). The evolving intellectual structure of the health informatics discipline: A multi-method investigation of a rapidly-growing scientific field. Available at SSRN 2498225.

Chuang, T. C., Liu, J. S., Lu, L. Y., & Lee, Y. (2014). The main paths of medical tourism: From transplantation to beautification. *Tourism Management, 45*, 49–58.

Chute, C. G., Cohn, S. P., & Campbell, J. R. (1998). A framework for comprehensive health terminology systems in the United States: Development guidelines, criteria for selection, and public policy implications. *Journal of the American Medical Informatics Association, 5*(6), 503–510.

Cimino, J. J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine, 37*(4–5), 394.

Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning* (pp. 233–240).

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics, 41*(6), 570–606.

Effken, J. A., & Carty, B. (2002). The era of patient safety: Implications for nursing informatics curricula. *Journal of the American Medical Informatics Association, 9*(Supplement 6), S120–S123.

Elmacioglu, E., & Lee, D. (2005). On six degrees of separation in DBLP-DB and more. *ACM SIGMOD Record, 34*(2), 33–40.

Epicoco, M., Oltra, V., & Saint Jean, M. (2014). Knowledge dynamics and sources of eco-innovation: Mapping the Green Chemistry community. *Technological Forecasting and Social Change, 81*, 388–402.

Evans, T. S., Calmon, L., & Vasiliauskaite, V. (2020). The longest path in the Price model. *Scientific Reports, 10*(1), 1–9.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics, 1*(4), 359–375.

Garfield, E., Pudovkin, A. I., & Istomin, V. S. (2003). Why do we need algorithmic historiography? *Journal of the Association for Information Science and Technology, 54*(5), 400–412.

Harris, M. R., Graves, J. R., Solbrig, H. R., Elkin, P. L., & Chute, C. G. (2000). Embedded structures and representation of nursing knowledge. *Journal of the American Medical Informatics Association, 7*(6), 539–549.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications, 124*, 226–251.

Hummon, N. P., & Dereian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks, 11*(1), 39–63.

Hung, S. C., Liu, J. S., Lu, L. Y., & Tseng, Y. C. (2014). Technological change in lithiumiron phosphate battery: The key-route main path analysis. *Scientometrics, 100*(1), 97–120.

Kim, M., Baek, I., & Song, M. (2018). Topic diffusion analysis of a weighted citation network in biomedical literature. *Journal of the Association for Information Science and Technology, 69*(2), 329–342.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM, 46*(5), 604–632.

Kuan, C. H. (2020). Regarding weight assignment algorithms of main path analysis and the conversion of arc weights to node weights. *Scientometrics, 124*(1), 775–782.

Lathabai, H. H., George, S., Prabhakaran, T., & Changat, M. (2018). An integrated approach to path analysis for weighted citation networks. *Scientometrics, 117*(3), 1871–1904.

Liu, J. S., & Kuan, C. H. (2016). A new approach for main path analysis: Decay in knowledge diffusion. *Journal of the Association for Information Science and Technology, 67*(2), 465–476.

Liu, J. S., & Lu, L. Y. (2012). An integrated approach for main path analysis: Development of the Hirsch index as an example. *Journal of the Association for Information Science and Technology, 63*(3), 528–542.

Liu, J. S., Lu, L. Y., & Ho, M. H. C. (2019). A few notes on main path analysis. *Scientometrics, 119*(1), 379–391.

Lu, L. Y., & Liu, J. S. (2013). An innovative approach to identify the knowledge diffusion path: The case of resource-based theory. *Scientometrics, 94*(1), 225–246.

Lucio-Arias, D., & Leydesdorff, L. (2008). Main-path analysis and path-dependent transitions in HistCite™-based historiograms. *Journal of the Association for Information Science and Technology, 59*(12), 1948–1962.

Martinelli, A. (2012). An emerging paradigm or just another trajectory? Understanding the nature of technological changes using engineering heuristics in the telecommunications switching industry. *Research Policy, 41*(2), 414–429.

Martinelli, A., & Nomaler, Ö. (2014). Measuring knowledge persistence: a genetic approach to patent citation networks. *Journal of Evolutionary Economics, 24*(3), 623–652.

McGuire, T. G. (1981). Price and membership in a prepaid group medical practice. *Medical Care, 19*(2), 172–183.

Mettler, T., & Raptis, D. A. (2012). What constitutes the field of health information systems? Fostering a systematic framework and research agenda. *Health Informatics Journal, 18*(2), 147–156.

Miller, R. A., Gardner, R. M., Johnson, K. B., & Hripcsak, G. (2005). Clinical decision support and electronic prescribing systems: A time for responsible thought and action. *Journal of the American Medical Informatics Association, 12*(4), 403–409.

Mills, R., Fetter, R. B., Riedel, D. C., & Averill, R. (1976). AUTOGRP: An interactive computer system for the analysis of health care data. *Medical Care, 14*(7), 603–615.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 262–272).

Mina, A., Ramlogan, R., Tampubolon, G., & Metcalfe, J. S. (2007). Mapping evolutionary trajectories: Applications to the growth and transformation of medical knowledge. *Research Policy, 36*(5), 789–806.

Nadri, H., Rahimi, B., Timpka, T., & Sedghi, S. (2017). The top 100 articles in the medical informatics: A bibliometric analysis. *Journal of medical systems, 19*(10), 150.

Park, H., & Magee, C. L. (2017). Tracing technological development trajectories: A genetic knowledge persistence-based main path approach. *PloS one, 12*(1), Article e0170895.

Pauker, S. G., Pauker, S. P., & McNeil, B. J. (1981). The effect of private attitudes on public policy: Prenatal screening for neural tube defects as a prototype. *Medical Decision Making, 1*(2), 103–114.
Penchansky, R., & Fox, D. (1970). Frequency of referral and patient characteristics in group practice. *Medical care, 8*(5), 368–385.
Raghupathi, W., & Nerur, S. (2008). Research themes and trends in health information systems. *Methods of Information in Medicine, 47*, 435–442.
Raghupathi, W., & Nerur, S. (2010). The intellectual structure of health and medical informatics. *International Journal of Healthcare Information Systems and Informatics (IJHISI), 5*(4), 20–34.
Sedgewick, R., & Wayne, K. (2011). *Algorithm* (4th Ed.). Princeton University. Addison-Wesley Available at http://algs4.cs.princeton.edu/ .
Sweeney, J. (2017). Healthcare informatics. *Online Journal of Nursing Informatics, 21*(1).
Tu, Y. N., & Hsu, S. L. (2016). Constructing conceptual trajectory maps to trace the development of research fields. *Journal of the Association for Information Science and Technology, 67*(8), 2016–2031.
Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in Complex Systems, 10*(01), 93–115.
Westberg, E. E., & Miller, R. A. (1999). The basis for using the Internet to support the information needs of primary care. *Journal of the American Medical Informatics Association, 6*(1), 6–25.
Woodward, C. A., McConvey, G. A., Neufeld, V., Norman, G. R., & Walsh, A. (1985). Measurement of physician performance by standardized patients: Refining techniques for undetected entry in physicians' offices. *Medical Care*, 1019–1027.
Xiao, Y., Lu, L. Y., Liu, J. S., & Zhou, Z. (2014). Knowledge diffusion path analysis of data quality literature: A main path analysis. *Journal of Informetrics, 8*(3), 594–605.
Yeo, W., Kim, S., Lee, J. M., & Kang, J. (2014). Aggregative and stochastic model of main path identification: A case study on graphene. *Scientometrics, 98*(1), 633–655.
Yu, D., & Sheng, L. (2020). Knowledge diffusion paths of blockchain domain: The main path analysis. *Scientometrics, 125*(1), 471–497.