Characteristic Analysis of Community Expansion and Convergence Patterns in Scientific Knowledge Network Diffusion: A Case Study in the Medical and Health Information Domain (Postprint)

Authors: Yue Lixin, Zhou Xiaoying, Liu Ziqiang

Date: 2023-04-01T16:15:57+00:00

Abstract

[Purpose/Significance] Knowledge units in scientific knowledge networks exhibit certain clustering and community characteristics. Revealing the fundamental patterns and features of community expansion and convergence during the temporal evolution of scientific knowledge network diffusion holds significance for expanding and deepening research on the laws of scientific knowledge diffusion and transmission. [Method/Process] First, an adjacency matrix is established based on citation relationships to construct a disciplinary knowledge network, and the Louvain community detection algorithm from complex network analysis is employed to partition the domain knowledge network into communities; then, network representation learning techniques are utilized to represent and compute the characteristics of community expansion and convergence; finally, using the time sequence as the logical thread, dynamic tracking and modeling of the expansion and convergence evolution processes of different communities are conducted, thereby revealing the fundamental patterns and features of community expansion and convergence during the temporal evolution of scientific knowledge networks. [Results/Conclusion] A case study is conducted in the field of health informatics. The research finds that the development trend of the community expansion pattern conforms to the Logistic model in S-shaped curve functions, while the development trend of the community convergence pattern conforms to the BiHill model in S-shaped curve functions.

Full Text

Preamble

Analysis of Characteristics of Community Expansion and Convergence Patterns in Scientific Knowledge Network Diffusion: A Case Study in Medical Health Information

Yue Lixin[1], Zhou Xiaoying[1], Liu Ziqiang[2],[3]

[1]School of Information Resources Management, Renmin University of China, Beijing 100872
[2]Chengdu Library of Chinese Academy of Sciences, Chengdu 610041
[3]Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract: [Purpose/Significance] Knowledge units in scientific knowledge networks exhibit certain clustering and community characteristics. Revealing the basic patterns and features of community expansion and convergence during the temporal evolution of scientific knowledge network diffusion holds significance for expanding and deepening research on the laws of scientific knowledge diffusion and transmission. [Method/Process] First, an adjacency matrix was established based on citation relationships to construct a disciplinary knowledge network, and the Louvain community detection algorithm from complex network analysis was employed to partition the domain knowledge network. Network representation learning techniques were then utilized to represent and calculate community expansion and convergence characteristics. Finally, using time series as the logical thread, the evolution processes of different communities' expansion and convergence were dynamically tracked and modeled to reveal the fundamental patterns and characteristics of community expansion and convergence during the temporal changes in scientific knowledge networks. [Result/Conclusion] A case study in the medical health information field demonstrates that the development trend of community expansion patterns conforms to the Logistic model within S-shaped curve functions, while the development trend of community convergence patterns conforms to the BiHill model within S-shaped curve functions.

Keywords: knowledge network; community detection; graph embedding; expansion model; convergence model

Classification Number: G251
DOI: 10.13266/j.issn.0252-3116.2020.14.007

In the era of data science, the continuous improvement of the scientific and technological innovation environment, the explosive growth of scientific literature, and the rapid development of digital publishing and related technologies have facilitated the global dissemination of scientific literature, accelerated worldwide scientific knowledge exchange, and further promoted the collaborative development of science and technology. Currently, both industry and academia attach great importance to the storage, mining, and utilization of massive scientific literature. However, regarding the implicit knowledge in scientific documents, "nodes" are represented by scientific literature or related concepts and entities (knowledge elements, knowledge units, etc.), while "edges" connect these "knowledge nodes" through certain relationships (co-occurrence, citation, causal, etc.). In summary, numerous scholars have explored objective scientific knowledge networks from quantitative, knowledge-based, and network perspectives, analyzing these networks to reveal the objective laws of knowledge diffusion and transmission phenomena.

Existing research primarily explores knowledge diffusion through citation relationships between scientific documents from a micro perspective. However, since knowledge units do not exist in completely isolated or free-floating states but rather exhibit clustering and grouping (community) characteristics based on relationships, examining knowledge diffusion solely through micro-level citation relationships cannot effectively reveal the temporal change mechanisms of clusters (communities) during the diffusion process. Therefore, research on knowledge diffusion at the community level requires further deepening. This study attempts to reveal the basic patterns and characteristics of community expansion and convergence in citation networks during scientific knowledge network diffusion, thereby expanding and deepening research on scientific knowledge network diffusion and transmission.

During scientific literature dissemination, documents serve as the primary carriers of knowledge, forming scientific knowledge networks through authors' citation and collaboration relationships. Over time, scientific literature and their interconnections continuously increase, facilitating knowledge transmission and diffusion. Exploring the basic processes of temporal changes in scientific literature networks and revealing the laws of scientific knowledge dissemination provide important reference bases for scientific and technological decision-making and promote technological innovation. In recent years, the diffusion and evolution of scientific knowledge networks have become a focal point for scholars in library and information science. "Knowledge nodes" are represented by scientific literature or related concepts and entities, while "edges" connect these nodes through relationships such as co-occurrence, citation, and causality. Scholars have quantitatively, knowledgeably, and network-wise explored objective scientific knowledge networks to study and reveal the objective laws of explicit and tacit knowledge diffusion and transmission.

This research is supported by the National Natural Science Foundation of China project "Research on Information Credibility and Quality Control of Medical Health Websites" (Project No.: 71473260) and the National Social Science Foundation of China project "National Health Promotion and Health Service Strategies in Healthy China Construction" (Project No. 16AZD021).

Author Biographies: Yue Lixin (ORCID: 0000-0002-7268-7871), PhD candidate; Zhou Xiaoying (ORCID: 0000-0002-9116-1525), Professor, PhD supervisor, Corresponding author, E-mail: xyz-ruc@qq.com; Liu Ziqiang (ORCID: 0000-0003-1814-8655), PhD candidate.

Received: February 12, 2020
Revised: April 23, 2020
Page Range: 63-73
Responsible Editor: Xu Jian

2 Literature Review

2.1 Scientific Knowledge Diffusion

Researchers currently utilize citation networks and collaboration networks as primary carriers for knowledge diffusion studies, exploring objective scientific knowledge network diffusion from quantitative, networked, and knowledge-based perspectives to reveal the objective laws of explicit and implicit knowledge diffusion and transmission, yielding numerous excellent research outcomes.

Qiu Junping et al. [1] (2014) constructed a citation network in the domestic knowledge mapping field, integrating and refining the citation network from four dimensions: journals, institutions, authors, and keywords, and introduced temporal dimensions to analyze knowledge diffusion and evolution. They discovered that domestic knowledge mapping research diffused from science and technology management to library and information science, and was subsequently applied to education and other disciplines. Li Gang et al. [2] (2017) introduced complex network and hypergraph mathematical theories to construct a knowledge diffusion evolution model in research collaboration supernetworks, exploring the evolution patterns and dynamic mechanisms of knowledge diffusion in research collaboration networks. By reproducing real network organizational knowledge dissemination behaviors, they revealed the dynamic relationships between different network structural features, node preference selection, knowledge growth and aging, knowledge diffusion pathways, and knowledge propagation processes. Yue Zenghui et al. [3] (2019) analyzed knowledge diffusion characteristics in disciplinary citation networks using the social network field as an example, specifically using literature citations as carriers of disciplinary knowledge transmission paths. They examined intermediary characteristics and middleman role features from three aspects: central tendency, dispersion degree, and distribution shape. Results indicated frequent interdisciplinary knowledge exchange activities in the social network field, with large fluctuation ranges in the quantitative characteristics of disciplinary knowledge diffusion, high dispersion degrees, and distributions mostly presenting as right-skewed peaked curves.

Research on the basic laws of scientific knowledge diffusion has always been an important concern for library and information science scholars. However, current research emphasizes knowledge diffusion exploration through micro-level citation relationships (knowledge diffusion speed, breadth, etc.), while studies on the temporal change mechanisms of clusters (communities) during knowledge diffusion require further deepening.

2.2 Self-Organization Mechanism of Scientific Knowledge Growth

Scientific knowledge growth has long been a significant research topic in information science and scientometrics. Current findings indicate that scientific knowledge growth across various disciplines exhibits self-organization characteristics over time. Representative achievements include E.C.M. Noyons et al. [4] (1998), who noted that during science and technology development, scientific cognitive systems drive dynamic knowledge growth and aging through self-organization. Regarding the essence of scientific knowledge growth, some scholars explore from sociological academic exchange perspectives, such as L. Leydesdorff et al. [5] (2003), who view scientific knowledge growth as an internal self-organization process of the scientific community, treating all disciplines as subsystems within a larger scientific system [6-7]. Drawing on system dynamics theory to study dynamic growth patterns, they demonstrate that scientific research activities accumulate knowledge through socialized academic exchanges. Other scholars approach from philosophy of science, such as K. Popper [8] (2007), who, building on classical empiricism, rationalism, and critical rationalism, proposed that science begins with problems and scientific theories are tentative answers to these problems. Scientific development is a chain-like process from problem to problem, acquiring new knowledge by modifying previous knowledge, giving scientific knowledge growth self-organizing characteristics. Still others examine from logical perspectives, such as Jing Jipeng and Ma Feicheng [9] (2009), who view science as an unstable, logically chaotic system that achieves stable, orderly construction through rational logical organization, promoting scientific knowledge growth through self-organization.

In information science, scholars primarily study scientific knowledge growth through scientific literature. Liu Zeyuan et al. [10] (2012) utilized co-citation analysis, co-word analysis, and visualization methods to analyze the dynamic evolution of scientific knowledge structures, noting that scientific knowledge units decompose and converge, disperse and recombine, evolve and sublimate, derive and transform during scientific development, forming a self-organizing system from simple to complex, low-level to high-level, and chaotic to orderly. Wan Hao [11] (2017) argued that with systems theory introduced to scientometrics, scientific knowledge growth has gained new interpretations within the systems theory framework, namely that stable, orderly network structures are obtained through logical rational organization during scientific knowledge growth.

The self-organizing characteristics of scientific knowledge growth at the micro level lead to the precipitation, free movement, recombination, and renewal of discrete knowledge units, forming scientific knowledge network structures. Since knowledge units do not exist in completely isolated or free-floating states but exhibit clustering and grouping (community) characteristics based on relationships, scientific knowledge network diffusion is accompanied by community structure expansion and convergence phenomena.

2.3 Knowledge Network Community Evolution

Knowledge units within disciplinary fields do not exist in completely isolated or free-floating states but exhibit clustering and grouping characteristics based on explicit and implicit relationships [12]. Some scholars have attempted to explore the evolution mechanisms of knowledge clustering and grouping during scientific knowledge growth [13].

The concept of complex network communities has effectively advanced research on knowledge network community evolution. M. Girvan et al. [14] (2002) proposed the concept of community structure, defining communities as subgraphs or subnetworks in complex networks characterized by dense internal connections and sparse inter-community connections. They identified network communities or network clusters as one of the most common and important topological properties of complex networks. Building on this, library and information science scholars have conducted extensive research on complex network community evolution. For example, M. A. Bettencourt et al. [15] (2009) revealed the temporal evolution of knowledge network communities based on indicators such as network density, diameter, and connectivity, suggesting that a field's development from emergence to maturity can be understood as a process where authors initially conduct discrete, isolated research before gradually integrating to form unified understanding. Bai Rujiang [16] (2013) and Wang Xiaoguang [17] (2013) analyzed knowledge network community evolution to uncover and reveal development trends in research themes. Teng Guangqing [18-19] (2018) conducted cross-replication analysis from multiple dimensions—frequency, correlation, quantity, scale, and temporal sequence—to study the growth of domain knowledge communities under social tagging Folksonomy knowledge organization modes, noting that revealing knowledge community growth patterns and laws helps expand domain knowledge organization perspectives and grasp knowledge development trajectories from the perspective of mutual promotion and interference between knowledge units.

Research on knowledge network community evolution represents an expansion and deepening of scientific knowledge growth law research under complex network thinking. Particularly in the data science era, with explosive growth in scientific literature and increasingly close knowledge connections, temporal sequence analysis of knowledge network structural relationships (network communities, clustering and grouping) better reveals the underlying laws of phenomena such as knowledge cross-fertilization, derivation, and fusion during evolution.

2.4 Diffusion of Scientific Knowledge in Scientific Communities

T.S. Kuhn proposed a scientific development model theory in *The Structure of Scientific Revolutions*, with paradigm as the core concept, viewing scientific development as an endless process of paradigm transitions. He defined scientific paradigm [20] as "the shared commitments of a community within a disciplinary field."

D.J. Price used the concept of "invisible college" to refer to informal academic groups (scientific communities) derived from formal academic organizations, exploring the relationship between scientific internal social structures and scientific knowledge growth—namely, the relationship between disciplinary social organization and knowledge growth. Using academic journals as an example, he studied the growth patterns of academic journals from 1650-1950, proposing the literature exponential growth law [21-22]. D. Crane built upon Kuhn's scientific development paradigm theory, scientific community doctrine, and Price's invisible college and quantitative research on scientific knowledge growth. Through analysis of citation relationships in academic papers, he examined connections between scientists, aiming to use observable data to illustrate various informal, non-fixed social connections among scientists and the existence of "invisible colleges" in various disciplinary fields. He also analyzed characteristics of scientific knowledge and scientific communities at different stages of scientific literature growth curves [23], as detailed in Table 1 .

Table 1 Characteristics of Scientific Knowledge and Scientific Communities at Different Literature Growth Stages

Stage	Scientific Knowledge Characteristics	Scientific Community (Invisible College) Characteristics
1	Emergence of new paradigm	Few community members; unclear community characteristics
2	Resolution of major problems	Community members increase rapidly; community formation evident
3	Increasing specialization	Community members continue increasing; community stabilizes
4	Decline of some research areas	Community members decrease; significant decline characteristics

Analysis of Table 1 reveals that scientific knowledge growth and diffusion are closely related to scientific communities. In Stage 1, new paradigms attract some scientists, but communities are not yet clearly formed. In Stage 2, as research deepens, scientific knowledge growth and diffusion attract numerous scientists forming scientific communities. In Stage 3, research becomes increasingly specialized and communities stabilize. In Stage 4, due to scientific development, some research areas decline, community members decrease, and scientific knowledge development shows significant decline characteristics. Crane's research on knowledge diffusion in scientific communities offers theoretical reference for current studies. Moreover, with recent developments in complex network analysis techniques and methods (such as community detection, clustering algorithms, and network representation learning) and the explosive growth of scientific literature data, research on the temporal evolution of expansion and convergence in scientific knowledge network communities now has solid foundations in theory, data, and methodology.

In summary, this study builds upon Kuhn's scientific development paradigm theory and scientific community doctrine, Price's quantitative research on scientific knowledge growth, and integrates Price's invisible college research. Guided by complex network thinking and from the perspective of scientific knowledge diffusion, this research explores clustering phenomena during knowledge diffusion. It dynamically tracks and models the temporal evolution of expansion and convergence in scientific knowledge network communities to reveal fundamental patterns and laws, aiming to provide beneficial research perspectives for knowledge diffusion studies.

3 Research Design

3.1 Research Data

This study uses literature data from the "medical health information" field indexed in the PubMed database. The specific retrieval strategy involves searching PubMed with "medical/health informatics/information" as title search terms, with no limitation on publication years. The search formula is: (((medical information[Title]) OR health information[Title]) OR health informatics[Title]) OR medical informatics[Title] AND "NIH grants"[Filter]. The search results serve as core literature for the domain, with PMIDs exported. Citation crawling was then performed using R programming to construct a citation network, yielding 5,643 documents published between 1981-2019. Based on these PMIDs, corresponding bibliographic data were retrieved and downloaded (online retrieval tool: https://www.ncbi.nlm.nih.gov/sites/batchentrez) and saved locally for subsequent research.

Given the large time span of the research data, to effectively reveal community expansion and convergence patterns during knowledge diffusion, the period was divided into eight time segments (corresponding to eight periods) from 1981-2019. Since the data spans nearly 40 years and literature quantities were relatively small in the early period (1981-2000), dividing periods strictly by equal natural years would create disproportionate differences between early and later periods, hindering subsequent model construction. Therefore, flexible time windows were employed to ensure relatively balanced literature quantities across periods while conforming to scientific literature growth laws (i.e., Price's literature exponential growth law), thereby ensuring scientific validity and effectiveness. The distribution of literature quantities across periods is shown in Figure 1 [Figure 1: see original paper].

3.2 Research Methods and Process

The fundamental objective of this study is to objectively measure and analyze the temporal changes of community expansion and convergence in scientific knowledge diffusion based on literature citation data, and to summarize the inherent patterns underlying these phenomena through mathematical modeling.

The research methodology and basic process are as follows: First, construct disciplinary knowledge networks based on literature citation relationships, employ the Louvain community detection algorithm from complex network analysis to partition domain knowledge networks, then utilize network representation learning technology (graph embedding) to represent and calculate community expansion and convergence characteristics. Using time series as the logical thread, the expansion and convergence evolution processes of communities in scientific knowledge networks are dynamically tracked and modeled to reveal fundamental patterns and laws. The main methodological processes are detailed below.

3.2.1 Scientific Knowledge Network Construction This study defines scientific knowledge networks as data structure models describing scientific literature (nodes) and their citation relationships (edges) interacting in chronological order. Visually, these are graphs Gt = (Vt, Et) formed by nodes (documents) and edges (citation relationships) that change over time, representing the set of scientific knowledge networks during any time period [0,n] ($0 \leq t \leq n$).

Scientific knowledge network construction is fundamental to subsequent research. To analyze community expansion and convergence patterns, dynamic analysis of scientific knowledge networks across different time windows is required. The construction process can be summarized in two steps:

(1) Scientific knowledge networks are growing networks. To dynamically track and model community expansion and convergence evolution, time periods must be divided annually: T = {t_1-t_2, t_2-t_3, ..., t -t }, where n < m.

(2) Based on the divided time periods T = {t_1-t_2, t_2-t_3, ..., t -t }, citation matrices are constructed for each period to establish period-specific scientific knowledge networks GT = {G_1, G_2, G_3, ..., G }, laying the data foundation for subsequent research.

Period-specific scientific knowledge networks are constructed using time-slicing methods (also called semi-cumulative scientific knowledge networks [24]), where citation networks consist of: (1) documents published within the period (citing literature), and (2) documents cited by these publications from both within and before the period. Analyzing fully cumulative citation networks would obscure current literature-revealed communities due to accumulated historical citation information. Therefore, constructing citation network slices for different periods effectively explores and reveals temporal community evolution during knowledge diffusion.

3.2.2 Community Detection Based on Louvain Algorithm This study employs the Louvain algorithm [25-26] for community detection. The Louvain algorithm is a modularity-based community discovery algorithm that can effectively detect hierarchical community structures. Modularity is the primary metric for evaluating community detection results, with larger values indicating better detection effectiveness.

The modularity concept was first proposed by M.E. Newman et al. [27] (2004) and has gained recognition among complex network scholars, forming the basis for numerous community detection algorithms, with the Louvain algorithm being a representative achievement. The Louvain algorithm performs well in both efficiency and effectiveness, offering advantages such as interpretability, unsupervised learning, and computational speed. Both the NetworkX toolkit in Python and the Gephi software for complex network analysis integrate Louvain algorithm functionality.

Specifically, the Louvain algorithm identifies knowledge communities within each time window through four steps: initial partition, modularity optimization, community aggregation, and community detection. The key step is modularity optimization, calculated using Formula (1) [25]:

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

Where Q represents modularity; Aij denotes the edge weight between nodes i and j in the complex network; ki and kj represent the sum of weights of all edges connected to nodes i and j; ci and cj indicate the communities of nodes i and j, with δ(ci, cj) = 1 when i and j belong to the same community and 0 otherwise; m is the total weight of all edges in the network. The maximum Q value is 1, with values closer to 1 indicating stronger community structure.

3.2.3 Community Expansion and Convergence Feature Representation and Calculation Based on Graph Embedding

Traditional social network analysis methods (degree, centrality, density, etc.) cannot effectively measure community expansion and convergence characteristics in citation networks. With rapid deep learning technology development, network representation learning methods (Graph Embedding Method, GEM), also called graph embedding (representing nodes as real-valued vectors while preserving network structure and inherent node attributes), represent the latest extension of embedding techniques developed from Word2vec [28-29]. These methods have been widely applied in recommendation systems and computational advertising in recent years, with DeepWalk, LINE, and Node2vec being representative algorithms.

This study attempts to use network representation learning technology to better measure community expansion and convergence characteristics, employing Stanford University's open-source Node2vec for feature representation and calculation. Node2vec is an algorithm that represents nodes in networks as real-valued vectors. Using deep learning principles, it maps each node to a K-dimensional real-valued vector through a three-layer neural network (input-hidden-output), converting relationships between any two network nodes into relationships between corresponding vectors. This facilitates computation and storage without manual feature extraction (self-adaptability). Node2vec defines an objective function f(u) to learn nodes' local neighbor structures, enabling similar features for nodes belonging to the same community or having similar structures. The objective function is shown in Formula (2) [30]:

$$f(u) = \max_f \sum_{u \in V} \log P(NS(u)|f(u)) \quad \text{Formula (2)}$$

$$P(NS(u)|f(u)) = \prod_{n_i \in NS(u)} P(n_i|f(u))$$

Where f(u) is the objective function mapping node u to an embedding vector; V represents the set of nodes in the network; S denotes the strategy for obtaining node neighbors; NS(u) represents the set of neighboring vertices sampled for node u through sampling strategy S. The Node2vec model algorithm [30] is illustrated in Figure 2 [Figure 2: see original paper].

This study primarily uses Python for community expansion and convergence feature representation and calculation. The processing steps involve: first, using the Node2vec model algorithm (https://github.com/aditya-grover/node2vec) to represent each node in the citation network as a computable K-dimensional vector; then, combining community partition results from the previous step, calculating distances between community nodes (vectors) to obtain each community's area size (using the maximum distance among community nodes as the diameter); finally, analyzing temporal changes in each community's area across different periods to characterize and measure community expansion and convergence during citation network temporal changes. The basic approach is summarized in Figure 3 [Figure 3: see original paper].

3.2.4 Community Expansion and Convergence Model Construction and Pattern Analysis Finally, using time series (divided time windows) as the logical thread, the expansion and convergence evolution processes of communities in scientific knowledge networks are dynamically tracked and modeled to analyze community expansion and convergence phenomena during network temporal diffusion and summarize their basic patterns. This involves two substeps:

(1) Before fitting community expansion and convergence temporal data, visualize the temporal data using line charts and scatter plots based on calculated community characteristics to analyze temporal change features and determine appropriate functional curves for fitting.

(2) Typical fitting function models include exponential functions, power functions, hyperbolic functions, and S-shaped curve functions, with basic formulas and function images shown in Figure 4 [Figure 4: see original paper].

Through the above processing and analysis, this study objectively measures community expansion and convergence temporal change phenomena in scientific knowledge diffusion, reveals knowledge diffusion laws at the meso level (community) based on citation relationships, and scientifically summarizes inherent patterns behind these phenomena through mathematical modeling, holding certain theoretical and practical significance for expanding and deepening scientific knowledge network diffusion research.

4 Results Analysis

4.1 Scientific Knowledge Network Construction and Community Detection

Following the proposed methodological process, initial citation networks for eight periods in the "medical health information (health informatics)" field were constructed. The Louvain algorithm was then applied for community detection, performing modular analysis on input citation data. The initial resolution parameter was set to the default value of 5, with fine-tuning based on actual partition results to improve detection accuracy. The community detection results for each period are shown in Figure 5 [Figure 5: see original paper], where node size is proportional to centrality and layout uses the Fruchterman-Reingold (FR) algorithm.

Figure 5 displays the initial citation networks for eight periods, revealing community structures through literature node connections and clustering tightness. Analysis shows that the medical health information field's initial citation networks continuously grow over time, with clear community structures emerging in each period and exhibiting certain community scale expansion and convergence phenomena during temporal evolution, warranting further analysis.

4.2 Community Expansion and Convergence Feature Calculation Results Based on Node2vec

Building upon community detection results and citation network data, the Node2vec model algorithm was used to extract features from each period's citation network, representing nodes as 128-dimensional real-valued vectors. Community node vectors for each period were obtained (see Table 2), where Community ID indicates each node's community, PMID represents the unique identifier in PubMed, and dimn denotes the node's n-dimensional real-valued vector.

Table 2 Community Node Vectors for Each Period (Partial Results)

Community ID	PMID	Dim1	Dim2	...	Dim128
...	...	...	...	...	...

The community node vectors for each period were then mapped to two-dimensional planes using T-SNE [31] (T-Distribution Stochastic Neighbor

Embedding) for visualization analysis (see Figure 6 [Figure 6: see original paper]). T-SNE is a machine learning method for dimensionality reduction that helps identify associated patterns, with the primary advantage of preserving local structure in high-dimensional data—points close in high-dimensional space remain close in low-dimensional projections.

Comparative analysis between Figure 6 and Figure 5 shows that community node clustering and distribution in Figure 6 are largely consistent with traditional community detection results. This demonstrates that representing community nodes as vectors through Node2vec effectively reveals community structures in citation networks for each period, validating the feasibility of measuring community scale through node vector calculations. Therefore, using community node vector results, distances between community nodes were calculated (specifically using Euclidean metric), with the maximum distance selected as the diameter to calculate each community's area size. Temporal data construction for community area sizes across different periods enables subsequent mathematical modeling to effectively measure community scale expansion and convergence.

4.3 Temporal Analysis of Community Expansion and Convergence in Medical Health Information

Combining full-period network node vector T-SNE visualization and community scale temporal change data (see Figure 7 [Figure 7: see original paper]), this study interprets development trends of the top 5 hotspot communities by overall scale: internet health information, health information behavior, electronic health information systems and technology, health information evaluation, and health management, laying the foundation for subsequent model construction and pattern analysis.

(1) **Internet Health Information**: Rapid internet technology development has made health information access more convenient and efficient, shifting public health information acquisition from relying solely on hospitals and doctors to internet platforms offering convenient access. This has promoted the emergence and development of health service platforms. However, the public knows little about factors affecting health information accessibility, facing issues such as poor information quality and inefficient access using primary search engines and simple search terms. Improving health information accessibility and retrieval efficiency remains a research priority. Additionally, public health information security and privacy protection in big data environments are currently key research focuses. Figure 7 shows this community exhibited stable development in the first five stages, rapid growth to peak in stage 6, and declining trends in stages 7-8, reflecting that internet health information represents a key and hotspot area in health information research.

(2) **Health Information Behavior**: Growing public demand for health information has diversified access channels, with internet and mobile tech-

nologies providing more convenient and effective methods. Sustained public attention to health information has prompted various health information behaviors, with information retrieval, acquisition behaviors and their influencing factors becoming research priorities. Figure 7 shows this community's temporal changes align with internet health information, demonstrating stable development followed by rapid growth and gradual decline, representing a key and hotspot area in health information research.

(3) **Electronic Health Information Systems and Technology**: Current international research on electronic health information systems and technology is relatively mature, featuring both conceptual models for system development processes that promote continuous improvement of health information technology, and practical platforms and systems. Figure 7 shows this community exhibited stable development before stage 6 and gradually increased thereafter, reflecting its emergence as a key research area in medical health information.

(4) **Health Information Evaluation**: In network environments, the public can access relevant health information conveniently and timely, but online resource quality varies significantly. Ensuring health information quality and achieving quality control have gradually become research hotspots, increasing related research. Current health information evaluation focuses primarily on influencing factors and evaluation system construction, with future research emphasizing technical methods for information assessment. Figure 7 shows this community exhibited stable trends before stage 7 and rapid development thereafter, indicating it will become a research hotspot in health information in coming years.

(5) **Health Management**: With lifestyle and health concept transformations in recent years, health management has demonstrated new characteristics under new health models, gradually emerging as a novel health service concept and approach. This has driven research on health management service systems, currently focusing on three aspects: (1) disease treatment-centered medical health management service systems; (2) health management service technologies; (3) health management system composition. Figure 7 shows this community's temporal changes are similar to health information, indicating potential to become a domain research hotspot.

4.4 Model Construction and Pattern Analysis of Scientific Knowledge Network Community Expansion and Convergence

Based on the above analysis, temporal data on community scale changes were fitted and modeled. Typical fitting function models include exponential, power, hyperbolic, and S-shaped curve functions. Based on observational data and Price's curve (literature growth law), the temporal change process of community scale in scientific knowledge networks better fits S-shaped curve functions (with representative Logistic and BiHill models). Therefore, this study used S-shaped curve functions to fit community expansion and convergence temporal data to construct mathematical models, with results shown in Figure 8 [Figure 8: see original paper], presenting model construction results for five hotspot communities (top 5 by overall scale): internet health information, health information behavior, and electronic health information systems and technology.

Figure 8 shows adjusted R-squared values above 0.9 for all five hotspot communities, indicating good model construction results. Adjusted R-squared reflects model fit quality (values closer to 1 indicate better fit; negative values indicate significant deviation).

Based on Figure 8 and the top 5 communities' expansion and convergence models in the medical health information field, patterns and basic characteristics of community expansion and convergence were summarized and analyzed, as shown in Table 3 .

Table 3 Community Expansion and Convergence Patterns and Basic Characteristics

Pattern	Basic Characteristics	Fitting Function	Model Formula	Lifecycle Stage
Expansion	Scientific paradigm emergence; few community members; rapid member increase; growth stage	S-curve Logistic	$y = A_1 - A_2/(1+(x/x_0)) + A_2$	Rapid growth phase
Convergence	Scientific paradigm focus and deepening; clear community groups; gradual member decrease; decline stage	S-curve BiHill	$y = A_1 + (A_2 - A_1)/(1+10\hat{}((\log x_{01} - x)p_1)) \times (1 - 1/(1+10\hat{}((\log x_{02} - x)p_2)))$	Decline phase

Analysis reveals that with continuous information technology advancement over recent decades, the medical health information field has developed well, with coexisting community expansion and convergence patterns. Internet health information, health information behavior, and electronic health information systems and technology represent core domain research areas with clear community characteristics and numerous members. Overall, the field is transitioning from expansion to convergence patterns, with development trends fitting S-shaped curve functions. Specifically, internet health information, health information behavior, and electronic health information systems and technology as core research areas are transitioning from expansion to convergence, fitting the BiHill model (past maximum, decreasing phase, corresponding to convergence pattern). Health information evaluation and health management as emerging communities show clear growth momentum with increasing members and expanding scale, fitting the Logistic model (rapid growth phase, corresponding to expansion pattern).

These conclusions indicate that temporal changes in community expansion and convergence during scientific knowledge network diffusion follow certain function models that can describe pattern characteristics: (1) Expansion pattern features paradigm emergence, few initial members, rapid member increase, and rapid growth stage, fitting the Logistic model; (2) Convergence pattern features focused and deepened internal paradigms, clear community groups, gradual member decrease, and decline stage, fitting the BiHill model.

This study has limitations. Focusing solely on the medical health information field may limit result accuracy. The research did not delve into specific textual content for community expansion and convergence pattern analysis during scientific knowledge diffusion. Additionally, different community partition algorithms might yield different conclusions. Future work will expand research data to include science, engineering, and humanities/social sciences fields, and attempt to study expansion and convergence issues at the specific textual content and research theme levels.

References

[1] Qiu Junping, Li Xiaotao. Research on knowledge diffusion based on citation network mining and temporal analysis[J]. Information Studies: Theory & Application, 2014(7): 5-10.

[2] Li Gang, Ba Zhichao. Research on knowledge diffusion evolution model in research collaboration supernetwork[J]. Journal of the China Society for Scientific and Technical Information, 2017(3): 58-68.

[3] Yue Zenghui, Xu Haiyun. Research on knowledge diffusion characteristics in disciplinary citation networks[J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(1): 5-16.

[4] Noyons ECM, Raan AFJV. Monitoring scientific developments from a dynamic perspective: self-organized structuring to map neural network research[J]. Journal of the Association for Information Science & Technology, 1998, 49(1): 68-81.

[5] Leydesdorff L. The challenges of scientometrics: the development, measurement, and self-organization of scientific communication[M]. Wu Yun, trans. Beijing: Scientific and Technical Documentation Press, 2003.

[6] Leydesdorff L, Cozzens S, Peter VDB. Tracking areas of strategic importance

using scientometric journal mappings[J]. Research policy, 1994, 23(2): 217-229.

[7] Leydesdorff L. Statistics for the dynamic analysis of scientometric data: the evolution of the sciences in terms of trajectories and regimes[J]. Scientometrics, 2013, 96(3): 731-741.

[8] Popper KR. The logic of scientific discovery[J]. Yinshan academic journal, 2005, 12(11): 53-54.

[9] Jing Jipeng, Ma Feicheng, Zhang Xiang. Intelligence science theory[M]. Beijing: Science Press, 2009.

[10] Liu Zeyuan. Crossing academic watersheds[M]. Beijing: People's Publishing House, 2012.

[11] Wan Hao. Research on the growth pattern of scientific knowledge scale—based on mathematical modeling and simulation[D]. Beijing: University of Chinese Academy of Sciences, 2017.

[12] Liu Ziqiang, Xu Haiyun, Luo Rui, et al. Research on identification method of science and technology interaction patterns based on topic correlation analysis[J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(10): 997-1011.

[13] An Ning, Teng Guangqing, Bai Shuchun, et al. Dynamic evolution analysis of domain knowledge clustering[J]. Library and Information Service, 2018, 62(10): 85-93.

[14] Girvan M, Newman MEJ. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12): 7821-7826.

[15] Bettencourt LMA, Kaiser DI, Kaur J. Scientific discovery and topological transitions in collaboration networks[J]. Journal of informetrics, 2009, 3(3): 210-221.

[16] Bai Rujiang, Leng Fuhai. Research on k-clique community knowledge innovation evolution method[J]. Library and Information Service, 2013, 57(17): 94-99.

[17] Wang Xiaoguang, Cheng Qikai. Visual analysis of disciplinary topic evolution based on NEViewer[J]. Journal of the China Society for Scientific and Technical Information, 2013, 32(9): 900-911.

[18] Teng Guangqing. Emergence of domain knowledge correlation relationships based on frequency evolution[J]. Journal of Library Science in China, 2018, 44(3): 79-95.

[19] Teng Guangqing. Association-driven domain knowledge community growth[J]. Journal of Library Science in China, 2017, 43(3): 58-71.

[20] Kuhn TS. The structure of scientific revolutions[M]. Chicago: University of Chicago Press, 1962.

[21] Price DJ. Science since babylon[M]. New Haven: Yale University Press, 1961.

[22] Price DJ. Little science, big science[M]. New York: Columbia University Press, 1963.

[23] Diana C. Invisible colleges—diffusion of knowledge in scientific communities[M]. Chicago: The University of Chicago Press, 1972.

[24] Luo Shuangling, Zhang Wenqi, Xia Haoxiang. Disciplinary topic evolution analysis based on semi-cumulative citation network community detection—taking the "cooperation evolution" field as an example[J]. Journal of Intelligence, 2017, 36(1): 100-110.

[25] Blondel VD, Guillaume JL, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008, 2008(10): P10008.

[26] Blondel VD. Louvain algorithm[EB/OL]. [2019-07-17]. https://perso.uclouvain.be/vincent.blondel/research

[27] Newman MEJ, Girvan M. Finding and evaluating community structure in networks[J]. Physical review, 2004, 69(2): 026113.

[28] Zhou Lian. Exploration of Word2vec working principle and application[J]. Sci-Tech Information Development & Economy, 2015, 25(2): 145-148.

[29] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems 26. Cambridge: Neural Information Processing Systems Foundation Inc., 2013: 3111-3119.

[30] Grover A, Leskovec J. Node2vec: scalable feature learning for networks[C]//Proceeding of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2016: 855-864.

[31] Laurens VDM, Hinton G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(11): 2579-2605.

Author Contributions

Yue Lixin: Conceptualization, framework design, writing and revision;
Zhou Xiaoying: Conceptualization, final manuscript revision;
Liu Ziqiang: Framework design, experimental design, and manuscript revision.

Analysis on the Characteristics of Community Expansion and Convergence Mode in the Diffusion of Scientific Knowledge Network——Take the Field of Medical Health Information as an Example

Yue Lixin[1] Zhou Xiaoying[1] Liu Ziqiang[2],[3]

[1]School of Information Resources Management, Renmin University of China, Beijing 100872
[2]Chengdu Library of Chinese Academy of Sciences, Chengdu 610041
[3]Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract: [Purpose/significance] Knowledge units in scientific knowledge networks show certain clustering and community characteristics. Revealing the basic patterns and rules of community expansion and convergence in the process of changing the time series of scientific knowledge networks has certain significance for expanding and deepening the research on the diffusion and transmission of scientific knowledge. [Method/process] Firstly, the adjacency matrix was built based on the citation relation, and then the subject knowledge network was constructed. The Louvain community detection algorithm in complex network analysis is used to divide the domain knowledge network into communities. Then, the Graph Embedding technique was used to represent and calculate the community expansion and convergence characteristics. Finally, the time series was used as the logical clue to dynamically track and model the process of expansion and convergence of different communities, so as to reveal the basic patterns and laws of community expansion and convergence in the process of time series change of scientific knowledge network. [Result/conclusion] A case study in the field of health information shows that the trend of community expansion conforms to the Logistic model in the S-shaped curve function and the trend of community convergence conforms to the BiHill model in the S-shaped curve function.

Keywords: knowledge network; community detection; graph embedding; expansion model; convergence model