Information Research - Vol. 31 No. iConf (2026)

Unraveling technology diffusion through dynamic network and multi-dimensional mechanism analysis: evidence from natural language processing

Yang Junhao, Xu Haiyun, Haunschild Robin, Li Shuying, Liu Chunjiang, and Yu XueLi

DOI: https://doi.org/10.47989/ir31iConf64140

Abstract

Introduction. Understanding factors influencing technology diffusion is vital for optimizing technological environments and fostering innovation. Existing studies often overlook temporal dependence and lack multidimensional mechanism analysis. This study addresses these gaps by introducing a dynamic network perspective to analyze technology diffusion.

Method. We developed a framework that integrates topic extraction with dynamic relationship modeling. Using patent data, BERTopic was applied to identify technological topics and construct cross - time-slice diffusion networks. Social network analysis captured evolutionary patterns, while the temporal exponential random graph model (TERGM) jointly examined endogenous network structures, actor - relation effects, and exogenous factors.

Analysis. The natural language processing field was selected as a case study. Diffusion dynamics and mechanism factors were investigated through quantitative modeling of temporal networks.

Results. The network has become more cohesive yet decentralized. Core nodes remain but their bridging role weakens. Reciprocity strongly promotes diffusion. Topic influence, novelty, and knowledge quality positively drive relationship formation, while knowledge breadth and depth affect only the sender effect.

Conclusions. This study integrates dynamic networks with multidimensional mechanism analysis, bridging gaps in temporal evolution and mechanism exploration, and providing a reusable framework and empirical reference for technology diffusion research.

Introduction

Technological innovation is the key to enhancing national competitiveness and promoting social development, and most innovations rely on the re-creation of existing technologies. Technology diffusion is not only an important path for technological learning and integration but also the foundation of innovation; efficient diffusion can promote knowledge dissemination and management, and thus has attracted wide attention from academia and industry. However, diffusion is not a linear process but a complex system involving multiple actors and multidimensional factors. Existing studies are mostly based on cross-sectional and single-dimensional perspectives, neglecting temporal evolution and multiple effects, and therefore fail to reveal the dynamic formation mechanism of relationships. To address this, this paper introduces a dynamic network perspective to construct a technology diffusion network and employs the temporal exponential random graph model (TERGM) to comprehensively examine the influence of both endogenous and exogenous mechanisms on relationship formation, thereby systematically clarifying diffusion mechanisms and providing theoretical and empirical evidence for optimizing the technological environment, improving diffusion efficiency, and promoting industrial upgrading.

Literature review

Research on the measurement of technology diffusion

Quantifying technology diffusion helps to grasp its dynamics and support decision-making. Since diffusion is difficult to observe directly, studies often describe it using proxy indicators. Among them, academic citation networks are the most used, as they present the collection, organization, and dissemination of knowledge, and reflect the flow from being cited to citing (Abramo et al., 2019; Xia et al., 2022; Zhuge, 2006). Knowledge networks constructed on this basis have been applied in cross-disciplinary and cross-national analyses: Yan (2013) proposed the '*disciplinary trade*' framework and validated it with Scopus (Yan, 2016); Abramo & D'Angelo (2018) revealed international differences from a geographical perspective and proposed the index of Balance of Knowledge Flows (BKF) indicator (Abramo et al., 2019). Compared with papers, patent citations more directly present technological inheritance and reuse, being structured and quantifiable, and thus have become a core carrier (Huang & Wang, 2011; Ye et al., 2015). Patent citations have therefore become a mainstream indicator in science and policy studies (Zhang et al., 2019).

Current research has shifted its scale from the national/regional level to organizations and topics: Ye et al. (2015), based on USPTO data, revealed cross-national diffusion and core–periphery structures; Chen et al. (2019) analysed cross-regional dynamics through university–enterprise cooperative patents; Choe et al. (2016) and Zhou et al. (2018) identified key organizations and explicit/implicit flows. Using '*topics*' as meso-level units can better reveal patterns (Liang et al., 2024): Zhang & Dong (2020) proposed an LDA–HMM to describe topic inflow/outflow; Suominen et al. (2017) employed LDA to identify emerging and declining fields in mobile communications.

In terms of research dimensions, the academic community has characterized diffusion from three aspects: knowledge flow, technological topics, and network structure. Lin et al. (2019) divided flows into absorption, value-added, diffusion, and conducted modeling; Wang et al. (2022) constructed indicators from the perspectives of intensity, speed, and breadth; Yang et al. (2021) and Wang et al. (2023) described the evolution of global and hierarchical networks using topology, centrality, structural entropy, and mutation functions.

As for foresight studies, Smojver et al. (2020) used co-citation networks and preferential attachment to predict life cycles and future diffusion; Wang et al. (2021) integrated direct citation–co-citation–coupling three-dimensional networks to predict potential flows in the field of gene editing. Overall, although carriers, scales, and dimensions have been continuously deepened, there is still room for improvement in revealing temporal evolution and multi-mechanism coupling within a unified framework.

Research on the influencing factors of technology diffusion

Existing studies have examined the role of multiple factors in technology diffusion from a network perspective. At the regional level, Obschonka et al. (2023) employed negative binomial regression to reveal the geographical, technological, and social media mediation effects of knowledge spillovers among U.S. cities, and further found that psychological openness promotes diffusion and absorption. Xu et al. (2024) combined Quadratic Assignment Procedure (QAP) regression to evaluate the impact of global hydropower industry competition/cooperation networks on diffusion, as well as the moderating effects of geographical distance and industry stage. From a multi-perspective standpoint, Qiao et al. (2019) applied four types of network models to demonstrate the interdependence among linkage structures, knowledge roles, and selection mechanisms. Regarding technological attributes, Sun et al. (2024), based on quantile regression of IoT patent citations, showed that technology convergence and scalability exhibit contextual dependence and distinct mechanisms.

Overall, the literature—based on papers, patents, or topics—has revealed diffusion patterns from disciplinary, regional, and organizational perspectives. However, dynamic network research with technological topics as units remains scarce, making it difficult to depict the temporal characteristics and intrinsic mechanisms of inter-topic diffusion; moreover, macro-level factors are hard to directly extrapolate to the micro-level. Therefore, it is necessary to systematically explore the influencing factors of diffusion from the perspectives of temporal evolution and multidimensional integration, in order to deepen the understanding of its essential mechanisms.

Research method

As shown in Figure 1, this study constructs a dynamic diffusion network of technological topics based on patents, and identifies the mechanisms of relationship formation by combining social network analysis with TERGM. First, data are obtained and pre-processed from incoPat, and topics are extracted in stages using BERTopic; cross-period networks are then constructed based on patent–topic citations. Second, the diffusion evolution is characterized from the perspectives of overall structure and node centrality. Finally, TERGM incorporates three categories of variables—endogenous structures, actor–relation effects, and exogenous contextual factors—for estimation, thereby comprehensively assessing the role of internal and external mechanisms in relationship formation.

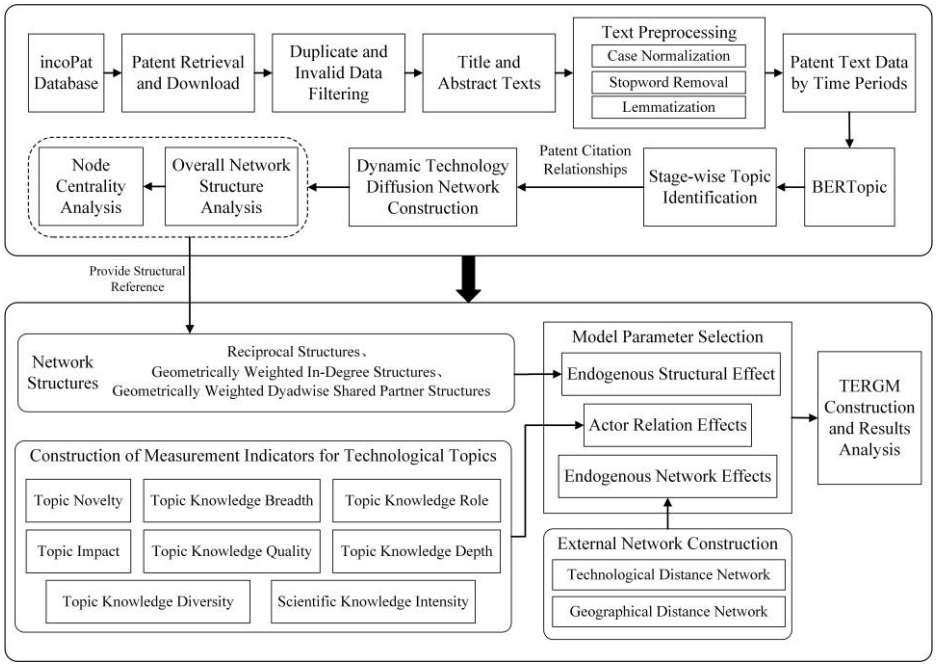


Figure 1 Research framework diagram

Construction of the dynamic technology diffusion network

Topic modeling method

In patent text topic modeling, text embeddings can overcome the limitations of methods like LDA that ignore semantic differences and contextual information (Farea et al., 2024). BERTopic, based on pretrained Transformers, generates embeddings, performs clustering, and then represents topics with intra-cluster TF-IDF, thereby balancing semantics and structure, enhancing interpretability, and achieving semantic disambiguation (Wang et al., 2024a). Accordingly, this study adopts BERTopic for modeling.

During the modeling process, several parameters were specified. In particular, the settings for the n-gram range, the number of topics, and the minimum topic size were informed by the studies of Kim et al. (2024) and Wang et al. (2024b). Random search was then employed to generate combinations of hyperparameters for iterative modeling. After each iteration, the information entropy of the topic word probability distributions was calculated; lower entropy indicates more coherent and well-defined topic semantics. The hyperparameter combination that yielded the lowest average entropy across all topics was selected as the optimal configuration, and this configuration was subsequently used for the final topic modeling of each stage.

Dynamic network construction method

After obtaining the topic modeling results, we construct a dynamic technology diffusion network with technological topics as nodes, distinguishing two types of relationships: (1) intra-period diffusion among topics, representing knowledge flows within a given time slice, and (2) cross-period continuity of the same topic, capturing its temporal evolution. Diffusion among topics is inferred from patent citation relationships: if any patent in Topic A cites a patent in Topic B, a citation link A→B is recorded, corresponding to technology diffusion from B to A. In the visualization, citations are shown as solid directed lines with thickness indicating intensity, while diffusion is represented by dashed lines in the opposite direction with the same intensity (Figure 2). Aggregating patents within each topic and their citations yields the topic-level diffusion network.

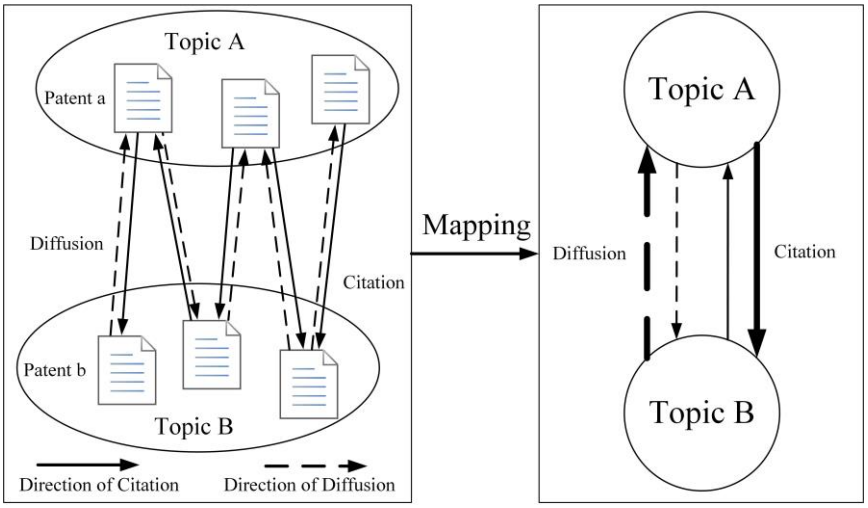


Figure 2 Schematic diagram of relationship mapping

To identify the '*same topic*' across periods, cosine similarity between topic embeddings of adjacent stages is calculated (Shen et al., 2024). This metric measures semantic correlation by the angle between vectors, with higher similarity indicating stronger correlation (Ran et al., 2024); the calculation is shown in Equation (1).

$$similarity(T_{k-1,i}, T_{k,j}) = cos(T_{k-1,i}, T_{k,j}) = \frac{T_{k-1,i} * T_{k,j}}{\|T_{k-1,i}\| * \|T_{k,j}\|} \quad (1)$$

Here, $T_{k-1,i}$ and $T_{k,j}$ represent the topic vectors in adjacent time stages k-1 and k, i and j represent the topics in these two stages. The cosine similarity ranges from zero to one, with higher values indicating closer semantic proximity between topics. In this study, the same topic is defined as follows: When the similarity is greater than 0.9 and the pair is each other's maximum similarity match across the two stages, they belong to the same topic.

Analysis of the characteristics of the technology diffusion network

To analyze the overall structural characteristics of the network, this study follows the methods of Choe et al. (2016) and Cao et al. (2022), selecting eight commonly used topological indicators for exploratory analysis: density, average degree, average path length, clustering coefficient, diameter, weighted average degree, and average clustering coefficient. Node centrality is an important indicator for measuring a node's influence and centrality within the network (Gong et al., 2022). Referring to Choe et al. (2016) and Cao et al. (2022), this study adopts Freeman's (1978) degree centrality, closeness centrality, and betweenness centrality to analyze the importance and value of network nodes.

Analysis of influencing factors of technology diffusion

Technology diffusion is jointly driven by exogenous conditions and endogenous dynamics, and the formation of network relationships is also influenced by various social processes. The TERGM can incorporate endogenous structural effects, actor–relation effects, and exogenous contextual factors to explore the formation mechanisms of technology diffusion relationships from a dynamic evolutionary perspective. Comprehensive consideration of network structure, technological topic characteristics, and external networks helps to more fully reveal the influencing factors of relationship formation and improve the model's goodness of fit and explanatory power.

Endogenous structural effects of the network

Endogenous structural effects describe the influence of network structure on edge formation and constitute a key mechanism in technology diffusion networks. This study focuses on three types of effects: reciprocity, convergence, and connectivity:

- Reciprocity refers to the tendency to form bidirectional connections in directed networks, reflecting '*feedback*' and basic reciprocal processes, and helps to explain the organizational principles of topology (Liu et al., 2021).
- Convergence measures the impact of indegree on the formation of new connections (in-K-star), reflecting preferential attachment and the '*Matthew effect*' of '*the strong becoming stronger*', which serves as an important endogenous driving force of network evolution (Liu et al., 2021; Yang et al., 2019).
- Connectivity corresponds to two-path structures, emphasizing the transmission of relationships through intermediate nodes and their mediating role in diffusion (Liu et al., 2021).

In directed weighted networks, geometrically weighted indegree (gwidegree) and geometrically weighted dyad shared partners (gwdsp) are used to measure convergence and connectivity, respectively. Geometrically weighted terms can enhance simulation convergence and estimation stability, avoid model degeneracy, and improve model fit (Hunter et al., 2008; Yang et al., 2018).

Actor–relation effects

Actor–relation effects refer to the influence of node attributes on edge formation; in this study, they correspond to the characteristics of technological topics, which affect diffusion relationships by altering their connection patterns. To test exogenous mechanisms, we incorporate in this paper multidimensional topic attributes into the model.

Four types of effects are examined: (i) homophily effect (topics with the same attributes are more likely to be connected, testing the homophilous linkage of knowledge roles); (ii) sender effect (topic attributes affect the probability of sending citations—technology inward absorption); (iii) receiver effect (topic attributes affect the probability of being cited by others—technology outward output) and (iv) difference effect (the influence of differences in continuous attributes on edge formation, focusing on knowledge diversity differences).

This study constructs eight technical topic–level indicators, which are incorporated into the four effects described above to investigate the roles of different topic attributes in the formation of technological diffusion relationships. The specific computational procedures for these indicators are described as follows:

（1）Novelty

Scientific and technological literature ages over time, and newly published literature is generally more novel. Topic novelty can be measured by calculating the average publication time of the documents it contains, taking into account both the number of documents and their temporal distribution (Zhu et al., 2024). The calculation method is shown in Equation (2).

$$novelty(T_i) = \frac{\sum_{p=1}^{x_i} y_p}{x_i} \tag{2}$$

Here, $novelty(T_i)$ denotes the novelty of topic i, x_i represents the number of patents contained in the topic, and y_p is the application year of the p^{th} document.

（2）Influence

The number of patent applications can reflect the degree of attention a topic receives; the larger the number, the greater the influence (Shen et al., 2024). In this study, topic influence is measured by the proportion of patents contained in a given topic relative to the total number of patents across all topics in that stage. The calculation of influence is shown in Equation (3).

$$influence(T_i) = \frac{x_i}{\sum_{i=1}^{M} x_i} \quad (3)$$

Here, $influence(T_i)$ denotes the influence of topic i, x_i represents the number of patents contained in that topic, and M denotes the total number of topics in that stage.

（3）Knowledge Breadth

To measure the knowledge breadth of a technological topic, the occurrences of IPC main group codes in its patents are first counted (if a patent contains n IPC codes, it is counted n times). Fine-grained IPC main group codes are selected to represent specific products, processes, and mechanisms (Zhang & Luo, 2020). Then, information entropy is used to measure breadth, as this method reflects both the number of categories and the uniformity of their distribution (Zeng et al., 2021). The calculation formula is shown in Equation (4).

$$H(X) = -\sum_{i=1}^{n} P(X_i)\, log_2\, P(X_i) \quad (4)$$

Here, $P(X_i)$ represents the proportion of occurrences of a specific IPC code within topic i relative to the total number of IPC codes in that topic. A higher information entropy value indicates greater knowledge breadth of the topic.

（4）Knowledge Depth

Knowledge depth reflects the degree of concentration of knowledge. Following previous studies (Cantwell & Piscitello, 2000; Li et al., 2021; Zhang & Baden-Fuller, 2010), this paper measures topic knowledge depth using a two-step method: In the first step, the Revealed Technological Advantage (RTA) is calculated, as shown in Equation (5):

$$RTA_{it} = \frac{\left(\frac{P_{it}}{\sum_t P_{it}}\right)}{\left(\frac{\sum_i P_{it}}{\sum_{it} P_{it}}\right)} \quad (5)$$

Here, P_{it} represents the number of patents in topic i that belong to technological category t; $\sum_t P_{it}$ is the total number of patents in topic i; $\sum_i P_{it}$ is the total number of patents belonging to category t across all topics; and $\sum_{it} P_{it}$ is the total number of patents across all topics. This indicator can be regarded as a comparative advantage index. When $RTA_{it} > 1$, it indicates that topic i has a relative advantage in category t. Its value is non-negative and has no upper bound.

In the second step, the coefficient of variation of topic i's RTA is calculated, as shown in Equation (6):

$$KDepth_i = \frac{\sigma_{RTAi}}{\mu_{RTAi}} \quad (6)$$

Here, μ_{RTAi} and σ_{RTAi} represent the mean and standard deviation, respectively, of the RTA values of topic i across all its patent technology categories. When a topic has relatively high advantages in only a few technological categories, it indicates greater knowledge depth; that is, the larger $KDepth_i$, the higher the knowledge depth of topic i. This measurement is consistent with knowledge breadth, as both are based on IPC codes at the main group level.

（5）Knowledge Roles

To measure the role and function of technological topics in the technology diffusion network, we follow the method of Choe et al. (2016), using the O–I index and betweenness centrality to classify the roles of nodes in the diffusion process. The outdegree of a node (the number of other nodes it cites) and its indegree (the number of times it is cited) can be used to calculate the O–I index, which reflects its knowledge output–input relationship. However, the O–I index cannot capture the overall importance

of a node in the network; betweenness centrality compensates for this limitation by evaluating the '*bridging*' role of the node. Based on this, the O–I index is used as the horizontal axis and betweenness centrality as the vertical axis to classify the knowledge roles of nodes.

In this paper, the calculation formula of the O–I index is improved based on the version proposed by Choe et al., as shown in Equation (7).

$$\text{O} - \text{I index} = \frac{(\text{Knowledge Outflow} - \text{Knowledge Inflow})}{(\text{Knowledge Outflow} + \text{Knowledge Inflow})} \tag{7}$$

Here, knowledge outflow refers to the weighted indegree of a node, while knowledge inflow refers to the weighted outdegree of a node. The O–I index ranges from −1 to 1. A value greater than 0 indicates that the node is cited more often than it cites others, and the larger the value, the more nodes cite it, implying higher knowledge quality. Conversely, a value less than 0 indicates that the node cites others more frequently, meaning it absorbs more external knowledge.

In this study, the knowledge roles of network nodes are divided into six categories:

(A) Knowledge-producing intermediaries (O–I > 0, high betweenness centrality);

(B) Knowledge-absorbing intermediaries (O–I < 0, high betweenness centrality);

(C) Knowledge-absorbing (O–I > 0, low betweenness centrality);

(D) Knowledge-producing (O–I < 0, low betweenness centrality);

(E) Knowledge-balancing intermediaries (O–I = 0, high betweenness centrality);

(F) Knowledge-balancing (O–I = 0, low betweenness centrality).

The median value (50th percentile) of node betweenness centrality is used as the threshold: Values above it are classified as '*high*', while values below or equal to it are classified as '*low*', in order to avoid classification bias (Zhou & Wei, 2018). In addition, isolated nodes also exist in the network.

（6）Knowledge Diversity

Diversity consists of three elements: variety, balance, and disparity (Stirling, 2007). The Rao–Stirling index is often used to measure interdisciplinarity, but it has limitations such as not satisfying balance monotonicity and difficulty in distinguishing between variety and balance (Rousseau, 2018). To overcome these shortcomings, it was later improved into DIV* by Rousseau (2019) and Leydesdorff et al. (2019). Compared with Rao–Stirling, DIV* is more effective in measuring disparity when it depends on the sample (Dong et al., 2024). Accordingly, we adopt DIV* as shown in Equation (8) to measure the cross-domain degree of topic patents, thereby characterizing their knowledge diversity.

$$DIV^{*} = n \cdot (1 - Gini) \cdot \sum_{t=1,s=1,t\neq s}^{t=n,s=n} \frac{d_{ts}}{n \cdot (n-1)} \tag{8}$$

Here, n denotes the number of IPC categories within a topic, and d_{ts} represents the technological distance between two IPC categories. Following previous studies, when the section, class, subclass, main group, or subgroup codes differ, d_{ts} is set to 16, 8, 4, 2, and 1, respectively, (Hou et al., 2024). The Gini coefficient is used to describe the inequality in the distribution among different IPC codes, and it is calculated according to Equation (9).

$$Gini = \frac{\sum_{q=1}^{n}(2q - n - 1)}{n\sum_{q=1}^{n} x_q} \tag{9}$$

Here, x_q denotes the number of patents contained in the q-th IPC category, arranged in ascending order by patent count.

（7）Knowledge Quality

This study uses the '*Incopat Value Index*' field of each patent document in the IncoPat patent database as an indicator to measure patent value, thereby reflecting the quality of the knowledge contained in the patents. To measure the knowledge quality of each topic, the calculation is shown in Equation (10):

$$KQuality = \frac{\sum_{v=1}^{10} w_v v}{N} \tag{10}$$

Here, v is the value of the Incopat Value Index, ranging from [1, 10]; w_v represents the number of patents within the topic that have a Incopat Value Index of v; N denotes the total number of patents contained in the topic. This formula can be used to measure the knowledge quality level of each topic.

（8）Scientific Knowledge Intensity

Topic scientific knowledge intensity is used to measure the extent to which scientific knowledge participates in the technological development process within a technological topic. Following the studies of Benson & Magee (2015), Papazoglou & Spanos (2018), and Wu, Xie, & Du (2024), this study calculates scientific knowledge intensity as the ratio of the number of scientific literature citations by patents within a topic to the sum of the number of scientific literature citations and the number of patents within that topic. The calculation is shown in Equation (11):

$$I_i = \frac{\sum_{p=1}^{x_i} \frac{S_{ip}}{P_{ip} + S_{ip}}}{x_i} \tag{11}$$

Here, x_i denotes the total number of patents contained in topic i; S_{ip} represents the number of scientific literature (scientific paper) references in the p^{th} patent document of topic i; and P_{ip} represents the number of patent references in the p^{th} patent document of topic i.

Exogenous contextual factors

The formation of network relationships is often symbiotic, with multiple associations simultaneously influencing technology diffusion. Accordingly, this study examines the effects of two types of external networks on edge formation: technological distance and geographical distance.

Technological distance: Cosine similarity is calculated based on the embedding vectors of topics in the same stage; the higher the similarity, the closer the semantics and the more similar the technologies. To measure differences, similarity is converted into technological distance, calculated according to Equation (12).

$$tech\_distance_{i,j} = 1 - similarity(T_{k,i}, T_{k,j}) \tag{12}$$

Here, $similarity(T_{k,i}, T_{k,j})$ denotes the cosine similarity between two topics. Based on this, the technological distance of all topic pairs is obtained using Equation (12), and a technological distance network is constructed: Topics are taken as nodes, and technological distance is used as edge weights, where a larger distance indicates a larger technological distance.

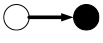
Geographical distance: First, geographical information is obtained based on the '*applicant country/region*' field of patents. The weight of each country within a given topic is determined by the ratio of the number of patents from that country to the total number of patents associated with the topic. This procedure is applied to all topics to construct a topic–country matrix, in which each row represents the distribution vector of a topic across countries. Based on these vectors, cosine similarity is employed to measure the geographical distribution similarity between any two topics i and j. To

capture the differences in their geographical distributions, the geographical distance between the two topics is defined as:

$$geo\_distance_{i,j} = 1 - similarity(i,j) \tag{13}$$

The closer the geographical distance is to 1, the greater the difference in geographical distribution between the two topics; the closer it is to 0, the more similar their distributions. Based on the geographical distances between topics within the same stage, a network is constructed in which topics are treated as nodes and geographical distances are used as edge weights.

Table 1 summarizes various effects and variables, and presents the corresponding network configurations (typical subgraphs) (Yang et al., 2019). On this basis, a TERGM is established and estimated using the R package tergm (Carnegie et al., 2015; Krivitsky et al., 2025; Krivitsky & Handcock, 2014; R Core Team, 2023.), to comprehensively test the effects of endogenous structures, actor–relation attributes, and exogenous contextual factors on the formation of diffusion ties. In the model, edges represent edge terms, equivalent to the regression intercept (Chen et al., 2023).

Category	Effect	Variable/Structure	Terms	Network Configuration	Explanation
Endogenous Structural Effects of the Network	Edge Effect	Edge	edges		Constant term, representing the baseline tendency of relationship formation, generally not interpreted
	Reciprocity Effect	Reciprocity Structure	mutual		Measures the tendency of bidirectional technology diffusion between technological topics
	Convergence Effect	Geometrically Weighted Indegree Structure	gwidegree		Reflects the distribution trend of the number of citation relationships received by technological topics
	Connectivity Effect	Geometrically Weighted Dyad Shared Partners Structure	gwdsp		Measures the tendency of technological topics to transmit diffusion relationships through multiple paths
	Homophily Effect	Topic Knowledge Role	nodematch		Measures whether technological topics with the same knowledge role in the network tend to mutually send diffusion relationships
Actor–Relation Effects	Sender Effect	Topic Influence	nodeocov		Measures whether technological topics with higher values of a certain attribute in the network tend to send citation relationships to other topics
		Topic Novelty			
		Topic Knowledge Quality			
		Topic Scientific Knowledge Intensity			

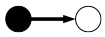
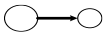
		Topic Knowledge Depth			
		Topic Knowledge Breadth			
	Receiver Effect	Topic Influence	nodeicov		Measures whether technological topics with higher values of a certain attribute in the network tend to receive citation relationships from other topics
		Topic Novelty			
		Topic Knowledge Quality			
		Topic Scientific Knowledge Intensity			
		Topic Knowledge Depth			
		Topic Knowledge Breadth			
	Difference Effect	Topic Knowledge Diversity	absdiff		Measures the impact of differences in topic knowledge diversity between network nodes on the formation of technology diffusion relationships
Exogenous Contextual Factors	Synergy Effect	Technological Distance	edgecov		Measures the impact of the technological distance and geographical distance on the formation of technology diffusion relationships
		Geographical Distance			

Table 1 Variable parameter settings and explanations

Empirical analysis

Data collection and processing

Natural language processing (NLP) is an important branch of artificial intelligence, integrating linguistics and computer science to enable machines to understand, generate, and utilize human language. Using NLP as an example, this study analyses the characteristics of its technology diffusion network and the factors influencing relationship formation. The data are obtained from incoPat, covering approximately 170 countries/regions/organizations and more than 200 million patents, including titles and abstracts in both Chinese and English, with patent families and citations processed. To ensure recall and precision, the search strategy was constructed with reference to prior studies and authoritative strategies (see Appendix, Table A1).

The retrieval date was May 5, 2024; invention patents and utility models filed between January 1,2000, and December 31, 2023, were selected, retaining only IPC sections G and H. After simple family consolidation, 129,182 patent families were obtained. Annual application volumes show: steady from 2000–2010, growth after 2010, acceleration from 2015, a slight decline in 2020, and significantly higher levels in the past decade compared with earlier years (see Figure 3).

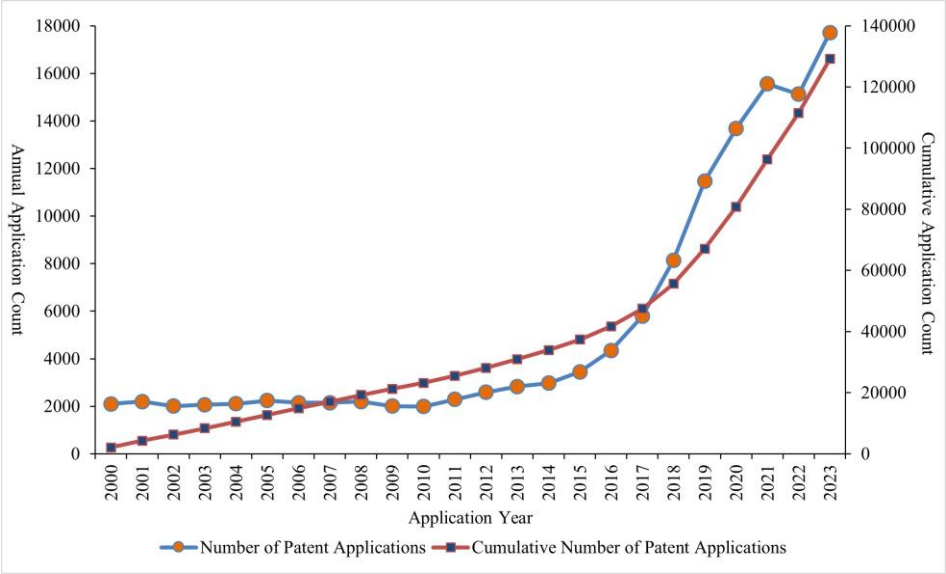


Figure 3 Number of NLP patent applications

For analytical convenience, the patent data are divided into five periods: 2000–2004, 2005–2009, 2010–2014, 2015–2019, and 2020–2023. The titles and abstracts of patents are merged as text data, and text preprocessing is conducted separately for each period, including the removal of non-alphabetic characters, case normalization, stop-word removal, and lemmatization, resulting in cleaned texts suitable for subsequent analysis.

Results of technology topic modeling

The minimum topic size was set to 0.5%–3% of the number of patents in each period, in order to balance reducing outlier topics and maintaining an interpretable number of topics (Kim et al., 2024). For each of the five-time stages, the model was iteratively trained 100 times, and the hyperparameter combination corresponding to the minimum average information entropy was selected as the final configuration for topic modeling. To facilitate understanding of topic content, following the method of Kim et al. (2024), ChatGPT -4) was employed to generate short labels based on topic words, and for the same topic, integrated labels and unified identifiers were summarized. Ultimately, topic identifiers and labels covering the entire time range were obtained (see Appendix, Tables A2).

Based on patent citations across five periods, we assign patents to topics and aggregate inter-topic citations to construct a directed, weighted topic citation network. Nodes represent technological topics, with size proportional to the number of patents, while directed weighted edges denote topic citations; edge direction reflects citing behavior (opposite to technology diffusion), and edge weight captures diffusion intensity. The networks for each period were visualized using Gephi (Figure 4). As shown in Figure 4, nodes are smaller and connections sparser in the first three periods, whereas node sizes increase and edges become denser in the last two periods, indicating a rising frequency of intra-field technology diffusion and a substantial expansion in the scale of technological topics over time.

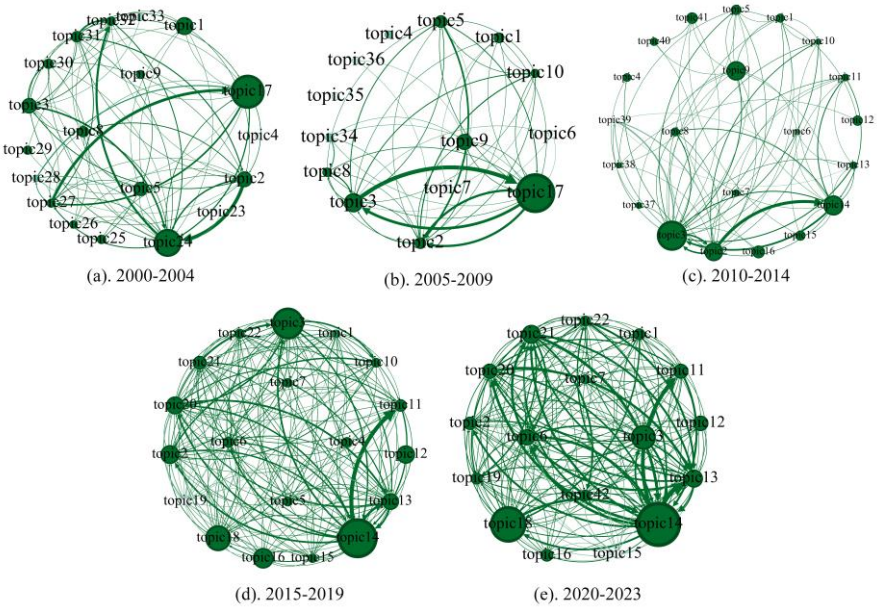


Figure. 4 Topic technology diffusion network in five periods

Dynamic characteristics of the technology diffusion network

Based on the constructed networks, Gephi was used to calculate the topological indicators for the five periods (Table 2). The results show that while network size fluctuated only slightly, the number of edges increased significantly, mainly concentrated in 2015–2019 and 2020–2023, indicating a rise in diffusion frequency in the later period, with a slight decline in 2005–2009. The network diameter decreased from 5 to 2, showing that the longest path was significantly shortened and efficiency improved; density increased simultaneously, suggesting closer connections among topics and more active diffusion.

Metric	2000-2004	2005-2009	2010-2014	2015-2019	2020-2023
Network Size	19	14	21	19	17
Number of Edges	76	49	77	191	185
Network Diameter	5	4	4	3	2
Network Density	0.222	0.269	0.183	0.558	0.68
Average Degree	4	3.5	3.667	10.053	10.882
Average Weighted Degree	14.579	33.571	17.238	76.632	119.471
Clustering Coefficient	0.379	0.539	0.477	0.71	0.758
Average Path Length	1.946	1.752	2.099	1.475	1.32

Table 2 Network topology metrics

The average degree and average weighted degree indicate that during 2005–2009, the diffusion scope was the smallest but the frequency was relatively high; in both 2015–2019 and 2020–2023, significant increases were observed, suggesting that diffusion scope and frequency grew simultaneously. During the same period, the network clustering coefficient rose and the average path length shortened, reflecting that both diffusion efficiency and frequency were significantly increased, and the overall trend became increasingly active.

To calculate node centrality indicators, the network data were imported into NetMiner4. The top five topics in degree centrality, closeness centrality, and betweenness centrality across the five time periods, along with their indicator results, are provided (see Appendix, Tables A3–A5), together with the mean centrality values of all nodes in each period.

Results of the analysis of factors influencing the formation of technology diffusion relationships

After measuring topic characteristics and constructing external networks, conditional maximum likelihood estimation (CMLE) was used to estimate the TERGM (see Table 3). The edge term of the baseline null model was significantly negative, consistent with the characteristics of real networks, indicating that diffusion relationships are non-random and that it is necessary to explore their influencing factors. The model settings are as follows: Model 1 includes only endogenous structural effects; Model 2 adds actor–relation effects; Model 3 incorporates both types of effects simultaneously; and Model 4 further adds exogenous contextual factors on this basis.

Variable Name		Null Model	Model 1	Model 2	Model 3	Model 4
Edge	edges	-2.543 *** (0.046)	-0.773 *** (0.090)	-10.918 *** (0.436)	-5.619 *** (0.628)	-5.698*** (0.646)
Reciprocity Effect	mutual		0.324 (0.185)		1.928 *** (0.240)	1.924 *** (0.243)
Convergence Effect	gwidegree		-0.618 (0.441)		-1.950 *** (0.575)	-1.943*** (0.566)
Connectivity Effect	gwdsp		-0.256 *** (0.017)		-0.239 *** (0.032)	-0.239*** (0.031)
Homophily Effect	Topic Knowledge Role			0.275 (0.196)	0.083 (0.177)	0.080 (0.177)
Sender Effect	Topic Influence			3.143 *** (0.365)	4.050 *** (0.438)	4.030*** (0.438)
	Topic Novelty			1.283 *** (0.350)	0.821 * (0.351)	0.811* (0.356)
	Topic Knowledge Quality			2.628 *** (0.519)	1.014 * (0.508)	1.014 * (0.514)
	Topic Scientific Knowledge Intensity			1.266 ** (0.413)	0.236 (0.432)	0.255 (0.435)
	Topic Knowledge Breadth			1.566 *** (0.394)	1.118 ** (0.369)	1.107 ** (0.380)
	Topic Knowledge Depth			1.856 *** (0.380)	1.903 *** (0.399)	1.917 *** (0.413)
Receiver Effect	Topic Influence			2.957 *** (0.367)	2.664 *** (0.432)	2.662 *** (0.422)
	Topic Novelty			1.655 *** (0.354)	1.003 ** (0.331)	0.991 ** (0.326)
	Topic Knowledge Quality			2.960 *** (0.522)	0.986 * (0.456)	0.992 * (0.463)
	Topic Scientific Knowledge Intensity			1.272 ** (0.415)	0.208 (0.372)	0.210 (0.372)
	Topic Knowledge Breadth			0.999 * (0.396)	0.486 (0.348)	0.482 (0.357)
	Topic Knowledge Depth			1.023 ** (0.380)	0.545 (0.370)	0.565 (0.375)
Difference Effect	Topic Knowledge Diversity			0.292 (0.365)	0.453 (0.275)	0.435 (0.284)
Synergy Effect	Technological Distance					0.125 (0.292)
	Geographical Distance					-0.036 (0.246)
AIC		3597.903	3468.873	1335.498	1127.277	1126.634
BIC		3604.741	3509.898	1438.061	1264.028	1277.059

Table 3 TERGM analysis results of technology diffusion network

Note: *, ** and *** indicate significance at $p < 0.1$, $p < 0.05$, and $p < 0.01$, respectively; values in parentheses represent standard deviations.

AIC/BIC results show that the null model performed the worst. After adding endogenous structural and actor–relation effects (Models 1–3), both indicators decreased significantly, with Model 3 reaching the lowest values. Model 4's AIC was slightly better than that of Model 3, but its BIC was slightly higher, indicating that exogenous contextual factors improved the fit but increased model complexity. Overall, Model 4 was selected.

Regarding endogenous structures, the reciprocity effect was significantly positive (1.924); the convergence effect (gwidegree) was significantly negative (–1.943, at the 1% level); the connectivity effect (gwdsp) was also significantly negative. These results suggest that decentralization has strengthened, while diffusion through intermediaries has weakened.

Regarding actor–relation effects, the homophily effect was not significant. On the sender side, influence, novelty, knowledge quality, breadth, and depth were all significantly positive (while scientific knowledge intensity was not significant), with influence and depth having the greatest effects. On the receiver side, influence, novelty, and quality were significantly positive (while scientific knowledge intensity, breadth, and depth were not significant). In the difference effect, differences in knowledge diversity were not significant.

Regarding exogenous contexts, both technological distance and geographical distance were not significant, possibly due to the modularity and transferability of NLP, as well as the openness of data and computing resources, which weaken geographical and technological barriers.

Conclusions

This study examines the characteristics and mechanisms of technology diffusion in natural language processing (NLP) from a dynamic network perspective. The results show that diffusion relationships have become increasingly close, balanced, and bidirectional over time. While core topics remain central, their bridging role has weakened, and direct diffusion has gained prominence.

Using the TERGM, we find that reciprocity significantly promotes relationship formation, whereas convergence and connectivity effects inhibit it. Topic influence, novelty, and knowledge quality drive diffusion on both the sending and receiving sides, while knowledge breadth and depth are significant only for senders. Technological and geographical distances are insignificant, suggesting a weakening of boundary constraints under digitalization and globalization.

Theoretically, this study shifts the analysis of technological diffusion determinants from traditional units of analysis to the topic level, emphasizing knowledge semantics. This perspective more directly captures knowledge transmission, recombination, and innovation processes. Building on this, we propose an integrated analytical framework that incorporates network structural effects, topic-level attributes, and external factors, allowing a unified assessment of endogenous and exogenous mechanisms and their relative influence on diffusion.

Methodologically, this study integrates BERTopic and TERGM into a unified framework covering topic extraction, network construction, and mechanism testing, overcoming the limitations of static analyses and enabling a more comprehensive depiction of dynamic diffusion processes.

This study relies exclusively on BERTopic for topic modeling, without comparison to alternative approaches (e.g., CTM, DTM, Top2Vec). Given differences across models in semantic representation, topic generation, and temporal dynamics, this may lead to incomplete topic identification. Future research could employ comparative or ensemble modeling strategies to enhance the robustness and completeness of topic modeling results.

Acknowledgements

This research was funded by the National Natural Science Foundation of China (No.72274113), Shandong Provincial Social Science Foundation (No.23CTQJ07), Shandong Provincial Natural Science Foundation (No. ZR202111130115), Beijing Natural Science Foundation (No.9242006) and the Taishan Scholar Foundation of Shandong province of China (tsqn202103069).

About the authors

Junhao Yang is a PhD candidate at the School of Information Management, Nanjing Agricultural University. His research interests focus on scientific and technical information analysis. He can be contacted at: yangjunhao163@gamil.com.

Haiyun Xu is a Professor at the School of Management, Shandong University of Technology. She received her Ph.D. from the University of Chinese Academy of Sciences. Her research interests include science, technology and industrial intelligence analysis. She can be reached at: xuhaiyunnemo@gmail.com.

Haunschild, Robin joined the Information Service for the institutes of the CPT section of the Max Planck Society in 2014. His current research interests include the advancement of bibliometrics and altmetrics as well as their application to specific fields of natural sciences, e. g., chemistry and climate change. He can be contacted at: R.Haunschild@fkf.mpg.de.

Shuying Li is an Associate Research Fellow at the Chengdu Library and Information Center, Chinese Academy of Sciences. Her research interests focus on patent analysis and bibliometrics. She can be contacted at: lisy@clas.ac.cn.

Chunjiang Liu is a Senior Engineer at the Chengdu Library and Information Center, Chinese Academy of Sciences. His research interests focus on scientific and technical literature mining. He can be contacted at: liucj@clas.ac.cn.

Xueli Yu is a masters' student at the School of Management, Shandong University of Technology. Her research interests focus on scientific and technical information analysis. She can be contacted at: 15063009091@163.com

References

Abramo, G., & D'Angelo, C. A. (2018). Who benefits from a country's scientific research? *Journal of Informetrics*, 12(1), 249–258. https://doi.org/10.1016/j.joi.2018.01.003

Abramo, G., D'Angelo, C. A., & Carloni, M. (2019). The balance of knowledge flows. *Journal of Informetrics*, 13(1), 1–9. https://doi.org/10.1016/j.joi.2018.11.001

Benson, C. L., & Magee, C. L. (2015). Quantitative determination of technological improvement from patent data. *PLOS ONE*, 10(4), e0121635. https://doi.org/10.1371/journal.pone.0121635

Cantwell, J., & Piscitello, L. (2000). Accumulating technological competence: Its changing impact on corporate diversification and internationalization. *Industrial and Corporate Change*, 9(1), 21–51. https://doi.org/10.1093/icc/9.1.21

Carnegie, N. B., Krivitsky, P. N., Hunter, D. R., & Goodreau, S. M. (2015). An Approximation Method for Improving Dynamic Network Model Fitting. Journal of Computational and Graphical Statistics, 24(2), 502–519. https://doi.org/10.1080/10618600.2014.903087

Cao, X., Zhu, J., & Yang, C. (2022). The "liquefaction" mechanism and empirical analysis of emerging technology innovation networks. *Science Research Management*, 43(2), 55–64. https://doi.org/10.19571/j.cnki.1000-2995.2022.02.007

Chen, M., Qin, Y., & Li, N. (2019). Dynamic evolution analysis of cross-regional knowledge flow and innovation cooperation networks. *Studies in Science of Science*, 37(12), 2252–2264. https://doi.org/10.16192/j.cnki.1003-2053.2019.12.015

Chen, Y., Wang, K., & Yu, C. (2023). Inter-provincial technology transfer in China: Spatial correlation and endogenous evolutionary mechanism. *Studies in Science of Science*, *41*(1), 38–50. https://doi.org/10.16192/j.cnki.1003-2053.20220426.003

Choe, H., Lee, D. H., Kim, H. D., & Seo, I. W. (2016). Structural properties and inter-organizational knowledge flows of patent citation network: The case of organic solar cells. *Renewable and Sustainable Energy Reviews*, *55*, 361–370. https://doi.org/10.1016/j.rser.2015.10.150

Dong, L., Fan, X., Li, Y., Wang, Z., Tao, Z., Zhang, H., Wei, R., & Li, Z. (2024). Research on the generality of major scientific infrastructure experimental stations and the correlation of innovation quality of their academic output. *Library and Information Service*, *68*(21), 107–119. https://doi.org/10.13266/j.issn.0252-3116.2024.21.010

Farea, A., Tripathi, S., Glazko, G., & Emmert-Streib, F. (2024). Investigating the optimal number of topics by advanced text-mining techniques: Sustainable energy research. *Engineering Applications of Artificial Intelligence*, *136*, 108877. https://doi.org/10.1016/j.engappai.2024.108877

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239. https://doi.org/10.1016/0378-8733(78)90021-7

Gong, Y., Wang, C., Wang, R., Xu, H., & Fang, S. (2022). Research on core patent identification from the perspective of complex networks. *Information Studies: Theory & Application*, *45*(10), 103–113. https://doi.org/10.16353/j.cnki.1000-7490.2022.10.014

Hou, J., Deng, X., & Tang, S. (2024). The impact of cross-domain knowledge integration on high-value patents. *Data Analysis and Knowledge Discovery*, 1–16. Advance online publication.

Huang, L., & Wang, N. (2011). A review of technology diffusion research from a patent perspective. *Science of Science and Management of S.&T.*, *32*(10), 27–34.

Hunter, D. R., Goodreau, S. M., & Handcock, M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association*, *103*(481), 248–258. https://doi.org/10.1198/016214507000000446

Kim, J., Jun, S., Jang, D., & Park, S. (2018). Sustainable technology analysis of artificial intelligence using Bayesian and social network models. *Sustainability*, *10*(1), Article 1. https://doi.org/10.3390/su10010115

Kim, K., Kogler, D. F., & Maliphol, S. (2024). Identifying interdisciplinary emergence in the science of science: Combination of network analysis and BERTopic. *Humanities and Social Sciences Communications*, *11*(1), 1–15. https://doi.org/10.1057/s41599-024-03044-y

Krivitsky, P. N., & Handcock, M. S. (2014). A separable model for dynamic networks. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1), 29–46. https://doi.org/10.1111/rssb.12014

Krivitsky, P. N., Handcock, M. S., Hunter, D. R., Goodreau, S. M., Morris, M., Carnegie, N. B., Butts, C. T., Leslie-Cook, A., Bender-deMoll, S., Wang, L., Li, K., Klumb, C., & Guillou, A. L. (2025). tergm: Fit, Simulate and Diagnose Models for Network Evolution Based on Exponential-Family Random Graph Models (Version 4.2.2) [Computer software]. https://cran.r-project.org/web/packages/tergm/index.html

Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2019). Diversity measurement: Steps towards the measurement of interdisciplinarity? *Journal of Informetrics*, *13*(3), 904–905. https://doi.org/10.1016/j.joi.2019.03.016

Li, Z., Sun, K., & Zhao, J. (2021). How does the knowledge base of enterprises regulate the performance of multi-source knowledge acquisition? —Threshold effect based on knowledge depth and breadth. *Studies in Science of Science*, 39(2), 303–312. https://doi.org/10.16192/j.cnki.1003-2053.20201119.003

Liang, S., Liu, X., & Chai, W. (2024). Discovery of important topics and knowledge flow paths from the perspective of topic-citation integration. *Data Analysis and Knowledge Discovery*, 8(2), 99–113.

Lin, B., Yang, X., & Hou, J. (2019). Comparative analysis of patent technology diffusion between China and the United States from the perspective of knowledge flow: A case study of carbon nanotube technology. *Journal of Intelligence*, 38(11), 43–49, 125.

Liu, L., Yan, X., Yang, L., & Song, M. (2021). The evolution and endogenous mechanism of international trade dependency networks. *China Industrial Economics*, 2, 98–116. https://doi.org/10.19581/j.cnki.ciejournal.2021.02.015

National Intellectual Property Administration. (2023). Notice of the Office of the National Intellectual Property Administration on the issuance of the Key Digital Technology Patent Classification System (2023). https://www.cnipa.gov.cn/art/2023/9/25/art_75_187769.html

Obschonka, M., Tavassoli, S., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2023). Innovation and inter-city knowledge spillovers: Social, geographical, and technological connectedness and psychological openness. *Research Policy*, 52(8), 104849. https://doi.org/10.1016/j.respol.2023.104849

Papazoglou, M. E., & Spanos, Y. E. (2018). Bridging distant technological domains: A longitudinal study of the determinants of breadth of innovation diffusion. *Research Policy*, 47(9), 1713–1728. https://doi.org/10.1016/j.respol.2018.06.006

PATENTSCOPE artificial intelligence index. (2019). Technology-trends. https://www.wipo.int/web/technology-trends/artificial_intelligence/patentscope

Qiao, T., Shan, W., Zhang, M., & Liu, C. (2019). How to facilitate knowledge diffusion in complex networks: The roles of network structure, knowledge role distribution and selection rule. *International Journal of Information Management*, 47, 152–167. https://doi.org/10.1016/j.ijinfomgt.2019.01.016

Ran, C., Tian, W., & Jia, Z. (2024). Patent technology topic evolution analysis method from the perspective of the technology life cycle: A case study of video image processing technology. *Information Studies: Theory & Application*, 47(9), 124–133. https://doi.org/10.16353/j.cnki.1000-7490.2024.09.013

R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/Rousseau, R. (2018). The repeat rate: From Hirschman to Stirling. *Scientometrics*, 116(1), 645–653. https://doi.org/10.1007/s11192-018-2724-8

Rousseau, R. (2019). On the Leydesdorff-Wagner-Bornmann proposal for diversity measurement. *Journal of Informetrics*, 13(3), 906-907. https://doi.org/10.1016/j.joi.2019.03.015

Schopf, T., Arabi, K., & Matthes, F. (2023). Exploring the landscape of natural language processing research (No. arXiv:2307.10652). *arXiv*. https://doi.org/10.48550/arXiv.2307.10652

Shen, J., Wang, W., Zhang, G., & Chen, H. (2024). Identification of emerging technology topics based on dynamic topic networks: A case study of the hydrogen fuel cell field. *Journal of Intelligence*, 43(9), 92–100.

Smojver, V., Štorga, M., & Zovak, G. (2020). Exploring knowledge flow within a technology domain by conducting a dynamic analysis of a patent co-citation network. *Journal of Knowledge Management*, 25(2), 433–453. https://doi.org/10.1108/JKM-01-2020-0079

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4(15), 707–719. https://doi.org/10.1098/rsif.2007.0213

Sun, B., Yang, X., Zhong, S., Tian, S., & Liang, T. (2024). How do technology convergence and expansibility affect information technology diffusion? Evidence from the internet of things technology in China. *Technological Forecasting and Social Change*, 203, 123374. https://doi.org/10.1016/j.techfore.2024.123374

Suominen, A., Toivanen, H., & Seppänen, M. (2017). Firms' knowledge profiles: Mapping patent data with unsupervised learning. *Technological Forecasting and Social Change*, 115, 131–142. https://doi.org/10.1016/j.techfore.2016.09.028

Wang, C., Xu, H., Wu, H., Qi, Y., & Chen, L. (2023). Research on the identification method of disruptive technology knowledge network diffusion characteristics based on dynamic structural entropy. *Library and Information Service*, 67(24), 54–71. https://doi.org/10.13266/j.issn.0252-3116.2023.24.006

Wang, F., Wang, X., Xu, S., Lu, W., & Song, Y. (2021). Potential knowledge flow detection based on a three-dimensional citation network: A case study of the gene editing field. *Journal of the China Society for Scientific and Technical Information*, 40(2), 184–193.

Wang, L., & Liu, X. (2022). Quantitative research and implementation of technology topic diffusion based on patent data. *Data Analysis and Knowledge Discovery*, 6(6), 1–10.

Wang, Y., Jiang, X., & Zheng, Y. (2024a). Topic mining and evolution analysis of scientific and technological reports based on the BERTopic model: A case study of the biotechnology field. *Information Science*, 1–14. Advance online publication.

Wang, Z., Chen, J, Chen, J., & Chen, H. (2024b). Identifying interdisciplinary topics and their evolution based on BERTopic. *Scientometrics*, 129(11), 7359–7384. https://doi.org/10.1007/s11192-023-04776-5

Wu, K., Xie, Z., & Du, J. T. (2024). Does science disrupt technology? Examining science intensity, novelty, and recency through patent-paper citations in the pharmaceutical field. *Scientometrics*, 129(9), 5469–5491. https://doi.org/10.1007/s11192-024-05126-9

Xia, H., Hu, Q., & Wang, Z. (2022). Main path analysis of knowledge flow based on citation importance. *Journal of the China Society for Scientific and Technical Information*, 41(5), 451–462.

Xu, G., Wang, Y., Wang, L., & Zhou, Y. (2024). How do competition and collaboration promote green technology diffusion? Evidence from the global hydropower industry. *Journal of Cleaner Production*, 478, 143890. https://doi.org/10.1016/j.jclepro.2024.143890

Yan, E. (2016). Disciplinary knowledge production and diffusion in science. *Journal of the Association for Information Science and Technology*, 67(9), 2223–2245. https://doi.org/10.1002/asi.23541

Yan, E., Ding, Y., Cronin, B., & Leydesdorff, L. (2013). A bird's-eye view of scientific trading: Dependency relations among fields of science. *Journal of Informetrics*, 7(2), 249–264. https://doi.org/10.1016/j.joi.2012.11.008

Yang, G., Chen, L., Zhang, J., & Li, G. (2019). An explanatory framework for the formation of patent citation relationships: A perspective of the exponential random graph model. *Library and Information Service*, 63(5), 100–109. https://doi.org/10.13266/j.issn.0252-3116.2019.05.012

Yang, G., Liu, T., Chen, L., & Zhang, J. (2018). Influencing factors of patent citation relationship formation based on the ERG model. *Science Research Management*, 39(11), 122–131. https://doi.org/10.19571/j.cnki.1000-2995.2018.11.014

Yang, W., Yu, X., Zhang, B., & Huang, Z. (2021). Mapping the landscape of international technology diffusion (1994–2017): Network analysis of transnational patents. *Journal of Technology Transfer*, 46(1), 138–171. https://doi.org/10.1007/s10961-019-09762-9

Ye, X., Zhang, J., Liu, Y., & Su, J. (2015). Study on the measurement of international knowledge flow based on the patent citation network. *International Journal of Technology Management*, 69(3–4), 229. https://doi.org/10.1504/IJTM.2015.072971

Zeng, B, Lyu, H., Zhao, Z., & Li, J. (2021). Exploring the direction and diversity of interdisciplinary knowledge diffusion: A case study of professor Zeyuan Liu's scientific publications. *Scientometrics*, 126(7), 6253–6272. https://doi.org/10.1007/s11192-021-03886-2

Zhang, G., Liu, L., & Wei, F. (2019). Key nodes mining in the inventor–author knowledge diffusion network. *Scientometrics*, 118(3), 721–735. https://doi.org/10.1007/s11192-019-03005-2

Zhang, J., & Baden-Fuller, C. (2010). The influence of technological knowledge base and organizational structure on technology collaboration. *Journal of Management Studies*, 47(4), 679–704. https://doi.org/10.1111/j.1467-6486.2009.00885.x

Zhang, R., & Dong, Q. (2020). Discovery of knowledge flow patterns based on the LDA-HMM model. *Information Science*, 38(6), 67–75. https://doi.org/10.13833/j.issn.1007-7634.2020.06.010

Zhang, Z., & Luo, T. (2020). Network capital, exploitative and exploratory innovations——From the perspective of network dynamics. *Technological Forecasting and Social Change*, 152, 119910. https://doi.org/10.1016/j.techfore.2020.119910

Zhu, X., Zhang, J., Li, W., Liu, X., & Geng, G. (2024). Emerging topic discovery in the field of artificial intelligence from the perspective of international top conferences. *Information Studies: Theory & Application*, 47(9), 147–155. https://doi.org/10.16353/j.cnki.1000-7490.2024.09.015

Zhuge, H. (2006). Discovery of knowledge flow in science. *Communications of the* ACM, 49(5), 101–107. https://doi.org/10.1145/1125944.1125948

Use of generative artificial intelligence

We employed ChatGPT for the following purposes: (1) translating parts of sections of the text into English, (2) proofreading and correcting grammatical errors. We evaluated the output by cross-referencing the translated and revised content with the original text to ensure accuracy,

consistency, and alignment with the intended meaning. Additionally, we reviewed the final version to confirm that all technical terms and concepts were appropriately conveyed.

Appendix tables

No.	Search Query
#1	TIAB= ((‘natural language processing’ OR NLP OR ‘natural language’ OR (natural* AND language*) OR languag* OR linguist* OR sentenc*) AND ((morpholog* OR morpheme* OR syntax* OR semantic* OR pragmatics* OR phonology* OR phonetics* OR lexic*) OR (multimodality* OR multimodal* OR text* OR image* OR audio* OR video*) OR ((dialog* OR talk OR conversation) OR model*))) OR ((voice* OR speech* OR acoustic* OR sound* OR audio* OR phonetic*) AND (natural* AND language*)) OR ((voice* OR speech* OR dialogu* OR conversat* OR speaking* OR language*) AND corpus*)
#2	TIAB=(((‘natural language processing’ OR NLP OR ‘semantic processing’ OR ‘machine translation’ OR ‘optical character reader’ OR ‘syntax analysis’ OR ‘word frequency’ OR ‘word segmentation’ OR ‘knowledge graph’ OR ‘natural language query’ OR ‘question answering’) OR (‘automatic translation’ OR ‘intelligent translation’ OR ‘language converter’ OR ‘literal translation’ OR ‘direct translation’ OR ‘word-for-word translation’ OR ‘rule-based machine translation’ OR ‘interlingual machine translation’ OR ‘neural machine translation’) OR (‘semantic understanding’ OR ‘semantic recognition’ OR ‘semantic analysis’ OR ‘semantic retrieval’ OR ‘semantic segmentation’ OR ‘semantic classification’ OR ‘semantic fusion’))) AND IPC=(G06F16* OR G06F40* OR G06K9* OR G06N3* OR G06N5* OR G06V30* OR G06F40/40 OR G06F40/42 OR G06F40/44 OR G06F40/45 OR G06F40/47 OR G06F40/49 OR G06F40/51 OR G06F40/53 OR G06F40/55 OR G06F40/56 OR G06F40/58 OR G06V20/40 OR G06V30/262)
#3	TIAB= ((‘natural language processing’ OR NLP OR ‘natural language’ OR languag* OR linguist* OR sentenc*) AND ((dialogue OR chatbot OR ‘personal assistant’ OR ‘question answering’) OR (‘information extraction’ OR ‘text mining’ OR ‘content extraction’) OR (‘machine translation’) OR (morpholog* OR ‘morphological analysis’) OR (‘natural language generation’ OR NLG) OR (stemming OR lemmatization OR lemmatisation) OR (semantics OR ‘semantic analysis’) OR (‘sentiment analys’ OR ‘rhetoric function’ OR ‘opinion mining’ OR ‘polarit classification’)))
#4	TIAB= ((‘natural language processing’ OR NLP OR ‘natural language’ OR (natural* AND language*) OR languag* OR linguist* OR sentenc*) AND ((multimodal* OR text OR visual OR speech OR audio OR ‘programming language’ OR ‘programming languages’ OR ‘structured data’) OR (‘natural language interface’ OR ‘question answering’ OR ‘dialogue system’ OR ‘conversational agent’ OR ‘dialogue systems’ OR ‘conversational agents’) OR (‘semantic text processing’ OR ‘discourse’ OR ‘pragmatics’ OR ‘text complexity’ OR ‘word sense disambiguation’ OR ‘language model’ OR ‘language models’ OR ‘representation learning’ OR ‘semantic similarity’ OR ‘semantic search’ OR ‘semantic parsing’ OR ‘knowledge representation’) OR (‘sentiment analysis’ OR ‘opinion mining’ OR ‘emotion analysis’ OR ‘emotions analysis’ OR ‘polarity analysis’ OR ‘aspect-based sentiment analysis’ OR ‘stylistic analysis’ OR ‘intent recognition’) OR (‘syntactic text processing’ OR ‘syntactic parsing’ OR ‘tagging’ OR ‘morphology’ OR ‘chunking’ OR ‘phonology’ OR ‘text segmentation’ OR ‘text error correction’ OR ‘typology’ OR ‘phonetics’ OR ‘text normalization’) OR (‘linguistic theories’ OR ‘psycholinguistics’ OR ‘cognitive modeling’) OR((responsible OR trustworthy OR ethical OR low-resource OR robustness OR explainability OR interpretability OR green OR sustainable OR linguistics OR cognitive) AND (‘natural language processing’ OR NLP)) OR (‘reasoning’ OR ‘textual inference’ OR ‘common-sense reasoning’ OR ‘argument mining’ OR ‘numerical reasoning’ OR ‘machine reading

	comprehension' OR 'knowledge graph reasoning' OR 'fact verification' OR 'claim verification') OR ('multilinguality' OR 'machine translation' OR 'code switching' OR 'cross-lingual transfer' OR 'typology') OR ('information retrieval' OR 'document retrieval' OR 'passage retrieval' OR 'indexing' OR 'semantic search' OR 'text classification') OR ('information extraction' OR 'text mining' OR 'text classification' OR 'topic modeling' OR 'text clustering' OR 'summarization' OR 'named entity recognition' OR 'coreference resolution' OR 'term extraction' OR 'relation extraction' OR 'open information extraction' OR 'event extraction') OR ('text generation' OR 'paraphrasing' OR 'question generation' OR 'dialogue-response generation' OR 'data-to-text generation' OR 'captioning' OR 'summarization' OR 'speech recognition' OR 'text style transfer' OR 'code generation')))
Final Search Query	#1 OR #2 OR #3 OR #4

Table A1. NLP patent search expressions

For patent search query #1, this paper refers to the construction of patent search strategies in the NLP field in the study by Kim et al. (2018). Based on the Patent Classification System for Key Digital Technologies (2023) compiled by the China National Intellectual Property Administration (CNIPA) (Notice of the CNIPA Office on the Issuance of the Patent Classification System for Key Digital Technologies (2023), 2023), the keywords and IPC codes related to natural language processing were used to construct patent search query #2. At the same time, key phrases related to the NLP field from the PATENTSCOPE Artificial Intelligence Index (World Intellectual Property Organization, WIPO, 2019) were used to construct patent search query #3. In addition, Schopf et al. (2023) systematically classified and analysed research papers in the ACL Anthology and developed a taxonomy of the NLP field, analyzing the latest developments in the field. Based on this taxonomy, patent search query #4 was constructed. Finally, the NLP patent search strategy of this study was formed as the combination of these four different queries.

Topic ID	Topic Label	Number of Patents
topic1	Multimedia Content Analysis and Subtitle Generation Methods	2655
topic2	Machine Translation and Language Processing Methods	4610
topic3	Speech and Voice Recognition Methods	9974
topic4	Power and Control Systems in Vehicles and Emergency Systems	936
topic5	Foreign Language Learning and Teaching Methods	1302
topic6	Medical Diagnosis and Patient Information Processing	2957
topic7	Sign Language Gesture Recognition and Translation	1308
topic8	Chinese Character Input Methods and Systems	722
topic9	Image Processing and Text Display Systems	2014
topic10	Speech Translation and Voice Recognition Systems	766
topic11	Sentiment and Emotion Analysis in Text and Public Opinion	2782
topic12	Educational Tools and Methods for Language Instruction and Display	3633
topic13	Deep Learning-Based Question Answering and Knowledge Base Methods	3926
topic14	Semantic Text Analysis and Representation Methods	13102
topic15	Database Query Processing and Data Analysis Methods	1242
topic16	Software Testing, Simulation, and Development Methods	3341
topic17	Programming Languages and Information Modeling	4822
topic18	Semantic Image Segmentation and Feature Detection Techniques	8855
topic19	Knowledge and Social-Based Recommendation Systems	1491
topic20	Natural Language Dialogue Systems and Interaction Modeling	3245
topic21	Knowledge Graph Construction and Entity Relationship Modeling	3681

topic22	Information Retrieval and Search Methods	1433
topic23	Optical Character Recognition (OCR) Technology	520
topic24	Text Retrieval and Analysis Methods	452
topic25	Print Data and Image Processing	1544
topic26	Email Communication Processing	2162
topic27	XML Document Structure and Schema Design	1864
topic28	Database Querying and SQL Models	1154
topic29	Voice Signal Processing and Circuit Devices	6799
topic30	Speech Synthesis and Rhythm Methods	1510
topic31	Voice-based Telephone Service and Communication	1170
topic32	Natural Language Dialogue and Conversation Systems	647
topic33	Speech Signal Processing and Codebook Methods	1987
topic34	Handheld Device Input and Text Disambiguation	4139
topic35	Wireless Navigation and Tourist Guide Systems	1946
topic36	Mobile Messaging and Short Message Service (SMS)	2581
topic37	Chinese Language Module for Remote Control Systems	8807
topic38	Audio Device Connectivity and Listening Experience	913
topic39	N-gram Language Model and Training	2794
topic40	Webpage Content and HTML Markup Processing	1076
topic41	Programming Language Compilation and Execution	1338
topic42	Knowledge-Based Fault Detection and Operation in Electric Power Grids	482

Table A2. Topic identification and topic labels

	2000-2004		2005-2009		2010-2014		2015-2019		2020-2023	
Rank	Indegree	Outdegree	Indegree	Outdegree	Indegree	Outdegree	Indegree	Outdegree	Indegree	Outdegree
1	topic24 0.667	topic24 0.556	topic2 0.692	topic17 0.615	topic3 0.700	topic3 0.550	topic3 0.889	topic14 0.833	topic14 1.000	topic14 1.000
2	topic2 0.500	topic17 0.444	topic17 0.615	topic3 0.538	topic2 0.450	topic14 0.450	topic14 0.889	topic3 0.778	topic3 0.938	topic18 0.875
3	topic3 0.389	topic3 0.389	topic3 0.462	topic2 0.462	topic14 0.400	topic2 0.400	topic13 0.778	topic20 0.722	topic18 0.875	topic13 0.875
4	topic1 0.333	topic31 0.389	topic1 0.385	topic5 0.385	topic11 0.300	topic9 0.400	topic20 0.778	topic2 0.722	topic20 0.875	topic3 0.813
5	topic17 0.333	topic2 0.333	topic5 0.385	topic10 0.385	topic5 0.250	topic1 0.250	topic2 0.722	topic13 0.667	topic13 0.875	topic21 0.813
6	topic27 0.333	topic27 0.333	topic8 0.308	topic8 0.308	topic10 0.200	topic13 0.250	topic11 0.722	topic18 0.667	topic21 0.813	topic20 0.750
7	topic31 0.333	topic32 0.333	topic9 0.308	topic34 0.308	topic6 0.150	topic5 0.200	topic18 0.722	topic21 0.667	topic6 0.750	topic15 0.750
8	topic32 0.333	topic8 0.278	topic10 0.308	topic9 0.231	topic8 0.150	topic8 0.200	topic1 0.722	topic16 0.667	topic11 0.750	topic11 0.688
9	topic8 0.222	topic30 0.278	topic7 0.154	topic1 0.154	topic9 0.150	topic10 0.150	topic19 0.667	topic1 0.611	topic2 0.750	topic1 0.688
10	topic30 0.222	topic5 0.222	topic6 0.077	topic6 0.154	topic1 0.150	topic39 0.150	topic6 0.611	topic19 0.611	topic19 0.688	topic22 0.688
Mean	0.222	0.222	0.269	0.269	0.183	0.183	0.558	0.558	0.680	0.680

Table A3. Degree centrality metrics results (Top 10)

	2000-2004		2005-2009		2010-2014		2015-2019		2020-2023	
Rank	Indegree	Outdegree	Indegree	Outdegree	Indegree	Outdegree	Indegree	Outdegree	Indegree	Outdegree
1	topic24 0.698	topic24 0.587	topic2 0.716	topic17 0.641	topic3 0.736	topic3 0.669	topic3 0.892	topic14 0.857	topic14 1.000	topic14 1.000
2	topic2 0.618	topic17 0.494	topic17 0.665	topic3 0.592	topic2 0.579	topic14 0.602	topic14 0.892	topic3 0.818	topic3 0.941	topic18 0.889
3	topic3 0.554	topic3 0.469	topic1 0.548	topic2 0.549	topic14 0.559	topic9 0.602	topic13 0.803	topic20 0.750	topic18 0.889	topic13 0.889
4	topic31	topic31	topic3	topic5	topic11	topic2	topic20	topic2	topic20	topic3

	0.535	0.469	0.548	0.513	0.523	0.531	0.803	0.750	0.889	0.842
5	topic32 0.535	topic27 0.469	topic5 0.517	topic8 0.481	topic5 0.506	topic8 0.501	topic2 0.765	topic13 0.720	topic13 0.889	topic21 0.842
6	topic27 0.518	topic2 0.447	topic8 0.490	topic10 0.481	topic10 0.491	topic1 0.488	topic11 0.765	topic22 0.720	topic21 0.842	topic20 0.800
7	topic1 0.502	topic32 0.447	topic9 0.490	topic34 0.427	topic8 0.476	topic37 0.488	topic18 0.765	topic21 0.720	topic6 0.800	topic15 0.800
8	topic17 0.487	topic8 0.427	topic10 0.490	topic9 0.405	topic39 0.450	topic5 0.475	topic1 0.765	topic16 0.720	topic11 0.800	topic11 0.762
9	topic8 0.472	topic30 0.427	topic34 0.388	topic6 0.405	topic13 0.450	topic39 0.463	topic19 0.730	topic18 0.692	topic2 0.800	topic1 0.762
10	topic30 0.472	topic28 0.391	topic7 0.372	topic36 0.405	topic9 0.438	topic13 0.463	topic6 0.698	topic1 0.692	topic19 0.762	topic22 0.762
Mean	0.372	0.382	0.398	0.398	0.428	0.424	0.681	0.652	0.777	0.771

Table A4. Closeness centrality metrics results (Top 10)

Rank	2000-2004	2005-2009	2010-2014	2015-2019	2020-2023
1	topic24 0.219	topic17 0.186	topic3 0.429	topic14 0.090	topic14 0.089
2	topic2 0.135	topic2 0.144	topic14 0.177	topic3 0.068	topic3 0.050
3	topic3 0.080	topic3 0.100	topic2 0.140	topic5 0.068	topic18 0.035
4	topic5 0.055	topic1 0.061	topic5 0.096	topic15 0.058	topic13 0.033
5	topic27 0.049	topic5 0.038	topic9 0.079	topic1 0.036	topic21 0.029
6	topic25 0.046	topic10 0.028	topic41 0.057	topic20 0.035	topic20 0.028
7	topic1 0.044	topic9 0.018	topic1 0.016	topic2 0.031	topic2 0.022
8	topic17 0.037	topic8 0.008	topic16 0.013	topic18 0.024	topic1 0.014
9	topic30 0.019	topic7 0.002	topic8 0.011	topic13 0.022	topic12 0.009
10	topic32 0.018	topic6 0.000	topic15 0.010	topic22 0.019	topic11 0.009
Mean	0.039	0.042	0.050	0.026	0.021

Table A5. Centrality measure results (Top 10)