Journal of Systems Science and Information
May, 2025, Vol. 13, No. 3, pp. 464–484
DOI: 10.12012/JSSI-2024-0072

The Empirical Study of Knowledge Diffusion Based on Citation Networks

Zhiyuan GE
School of Economics and Management, Beijing University of Technology, Beijing 100124*, China*

Kanran LI
School of Mathematics and Science, Beijing University of Technology, Beijing 100124*, China*

Abstract This paper investigates the influence of various knowledge roles on knowledge diffusion empirically. Exponential random graph models (ERGM) are constructed, which provides a novel perspective for examining the factors that influence knowledge diffusion. Our empirical findings reveal that the endogenous structural effects of the network have a significant impact on the formation of diffusion relationships in citation networks and that there is a correlation between the number of the three knowledge roles - contributors, seekers and brokers - and the likelihood of citation relationship formation in citation networks.

Keywords knowledge diffusion; exponential random graph model; citation networks; knowledge roles

1 Introduction

Knowledge tends to be diffused and exchanged among different entities [1]. Citation relationships between papers, patents, authors, journals and institutions are the main research component of knowledge diffusion. Analysing the knowledge diffusion within the citation network is of both academic interests and practical significance for promoting knowledge production and dissemination. This paper focuses on this topic.

The factors influencing knowledge diffusion have received attention from scholars. Some scholars are concerned with aspects such as the individual's level of activity [2], innovativeness [3,4], factors of their social capital including collaborative experience, trustworthiness, and similarity in career length, and individual's position in the scientific publishing network [5]. On the macro-levels, scholars have focused on the size and structure of network organisations [6,7], the distribution of knowledge roles, reputation mechanisms in knowledge diffusion networks [8],

Received June 27, 2024, accepted March 6, 2025
Supported by R&D Program of Beijing Municipal Education Commission (KZ202210005013); Sichuan Social Science Planning Project (SC22B151)

and the incentives of the environment in which diffusion subjects participate in knowledge diffusion[9], as well as the knowledge competence of organisational culture[3].

The subjects involved in knowledge diffusion are assigned four different knowledge roles: Knowledge seekers, brokers, knowledge contributors, and lurkers, based on their different behavioral characteristics. The literature has demonstrated that all four knowledge roles are important for knowledge diffusion[10–14], but empirical understanding on the impact of knowledge roles on knowledge diffusion is still limited. In addition, there are studies on knowledge seekers and knowledge contributors[11,13–15]. Moreover, current research on knowledge diffusion fails to take into account both micro-attributes and macro-factors, and most of the current studies on knowledge diffusion factors are through simulation experiments relying on the setting of network-related attribute parameters[6,8,15]. The majority of the current research on the factors that influence knowledge diffusion is conducted through simulation experiments that rely on the establishment of network-related attribute parameters, as tracing the path of knowledge diffusion is an extremely challenging problem[16], In general, empirical research on the impact of knowledge roles, network node features, and network results during the dissemination process is lacking in present study on academic knowledge diffusion based on citation networks. Simulation models cannot grasp the diffusion issue in real-world networks even if they can investigate the mechanism and growth of knowledge diffusion. Empirical research is necessary to comprehend the genuine influence of the entities involved in knowledge diffusion on diffusion in actual networks, as simulation methods are incapable of achieving this.

The citation network is a typical knowledge diffusion network that represents the process of knowledge diffusion from the cited paper to subsequent citing papers[16]. Consequently, the process of knowledge diffusion can be reflected in the discussion of the citation network's knowledge diffusion. A research method known as the Exponential Random Graph Model (ERGM) is used to analyze the global structure of a network and infer the relationship prediction of nodes at the micro level. It is frequently employed in social network analysis to investigate the processes of information dissemination and social relationships. Consequently, we regard the citation network as the research object and employ the ERGM method to empirically analyze the subjects involved in knowledge diffusion in the knowledge network. This method is intended to evaluate the actual influence of various subjects on the knowledge diffusion process in real knowledge networks.

The structure of this paper is as follows. In Section 2 we present the relevant literature, and in Section 3, we propose an empirical research framework based on ERGM, and define knowledge roles within citation networks and formulate empirical hypotheses concerning knowledge roles in the dissemination of knowledge. In Section 4, we construct an empirical model for knowledge diffusion in citation networks and analyzed the empirical results using the R language toolkit. In Section 5, we discuss the results of empirical analysis and compared them with the research findings of other researchers based on simulation methods, as well as the future work of this paper. Section 6 concludes the article.

2 Related Work

Knowledge diffusion is the movement of knowledge through interconnected social and organizational networks, where nodes represent knowledge holders and edges represent transmission pathways[17]. Cowan and Jonard[18] proposed that an agent's reputation could affect the extent of knowledge diffusion. Qiao et al.[15] suggested that the macroscopic pattern of knowledge diffusion is much more than a simple aggregation of individual attributes, and due to the unbridgeable distance between micro- and macro-level studies of knowledge diffusion, the macroscopic pattern of knowledge diffusion cannot be deduced from local knowledge-seeking and sharing interactions among network members. Kim and Park[6], Cowan and Jonard[18] and several other scholars have pointed out in their studies that the topology of the network is an important factor influencing knowledge diffusion. Singh[19], Abramo et al.[20] and others have also pointed out that geographic location also has an impact on the knowledge diffusion process. Havakhor et al.[8], Qiao et al.[15] found that the distribution of knowledge roles, the interaction of selection mechanisms with the distribution of knowledge roles all affect knowledge diffusion in the network.

Welser et al.[21] characterized the behavioral characteristics of participants based on the communication features between network members, and divided the knowledge diffusion roles in online forums into four types: Contributors, seekers, brokers, and lurkers. Nam et al.[11], Marett and Joshi[22], Nonnecke and Preece[23], Gray[24], Ridings et al.[25], Cassi et al.[26], have demonstrated that the existence of four knowledge roles in knowledge diffusion networks, namely, seekers, brokers, contributors, and lurkers.

The definition of knowledge roles can be distinguished by the characteristics of communication among network members[21]. The main behaviour of knowledge seekers is to seek help, knowledge contributors primarily provide help and share knowledge, and brokers are intermediaries who facilitate knowledge flow between different communities or groups[15,21]. Studies by Welser et al.[21] defined lurkers as members of a network who had no or very few posts over a given period. It is evident that the behavioral characteristics of lurkers emphasize "relative silence", and whether "silence" means never posting, not posting for a certain period of time, or posting within a certain range, depends on the specific context.

In terms of the degree of influence of knowledge roles Qiao et al.[15] indicated that knowledge seekers exert more influence on knowledge diffusion than knowledge contributors. Nam et al.[11] also revealed that active knowledge seekers cannot solely contribute to knowledge dissemination alone, but high-quality knowledge contributors may benefit a wider audience. These studies demonstrated that the four knowledge roles have distinct functions in the knowledge diffusion process, exhibiting heterogeneity. However, the quantitative analysis of the impact of these four knowledge roles on knowledge diffusion remains unexplored in the existing literature.

As a complicated, dynamic, and emergent process, knowledge diffusion can be better understood by using simulation-based quantitative analyses to capture its dynamic variations[15]. Scholars have shifted their focus to simulation-based quantitative analysis in an attempt to obtain more legitimate research conclusions using simulation experiments, as there are concerns regarding the validity of qualitative research methodologies. Cowan and Jonard[18] modelled the effect of network structure on knowledge diffusion as a bartering process. Kim and Park[6] concluded that small-world networks are the most efficient and fair structure to achieve effective knowledge diffusion through simulation experiments. Havakhor et al.[8] also investigated the effects of reputation machanisims and distribution of knowledge roles based on simulation experiments.

Simulation based experimental research can only simulate knowledge diffusion mechanisms and cannot quantitatively analyze knowledge dissemination in real networks. Macy and Willer[27] pointed out that agent-based simulation methods simplify the intricate micro-angle behaviour and produce distortion to a certain extent in their research. Qiao et al.[15] also noted the drawbacks of simulation techniques that ignore the dimensions and complexity of knowledge.

To further investigate and identify the elements influencing information spread, empirical study is essential. The same concern was also raised by Kim and Park[6] in the outlook section of their study. Singh[19] has empirically evaluated the factors influencing the probability of generating knowledge flows using patent data and a regression framework. In their investigations into the variables influencing the knowledge diffusion, Wang and Zhang[28] also employed empirical methodologies.

3 Methodology

Using ERGM a researcher can identify which characteristics belong to network members in an observation network. Specifically, ERGM conciders both micro-node attributes and macro-attributes of an observation network. It can be utilized to determine whether the establishment of an observation network is influenced by specific attribute characteristics of its members (such as age, occupation, etc.) or by the relational model in the network formation process, aligning with our research objective of examining the effect of individual node attributes (pertaining to their knowledge roles) on knowledge diffusion. The study by Jiang and Chen[2] demonstrated that the ERGM is applicable for studying knowledge diffusion. Therefore, we consider using ERGM to empirically analyze the impact of different knowledge roles on knowledge diffusion in the network.

3.1 Study Design

We hypothesize that the endogenous structure of the citation network, the attributes of the nodes in the network (specifically, researchers' knowledge roles), and the interactions between the nodes will influence the degree of knowledge diffusion. We construct an empirical model utilizing exponential random graphs to analyze the degree of the impact of knowledge roles on knowledge diffusion. The empirical research framework is illustrated in Figure 1.

The research framework described above is divided into three parts. The first part is to construct a citation network, define different knowledge roles in the network, and obtain a knowledge diffusion network based on citation relationships. Then, by identifying the endoge-

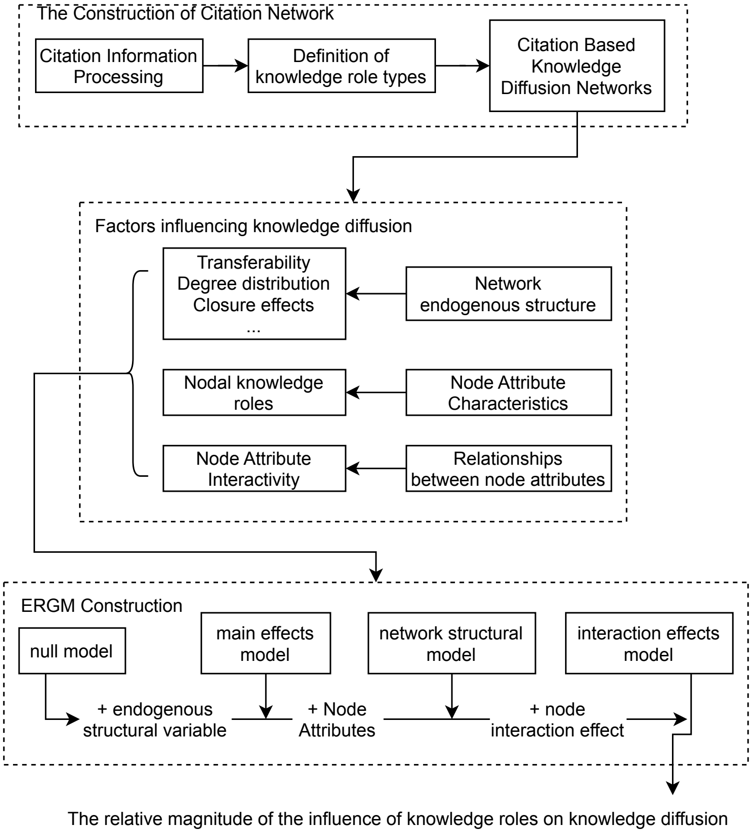

The Construction of Citation Network
Citation Information Processing
Definition of knowledge role types
Citation Based Knowledge Diffusion Networks
Factors influencing knowledge diffusion
Transferability Degree distribution Closure effects ...
Network endogenous structure
Nodal knowledge roles
Node Attribute Characteristics
Node Attribute Interactivity
Relationships between node attributes
ERGM Construction
null model
main effects model
network structural model
interaction effects model
+ endogenous structural variable
+ Node Attributes
+ node interaction effect
The relative magnitude of the influence of knowledge roles on knowledge diffusion

Figure 1 Research framework on factors influencing knowledge diffusion

nous structure of the network, node attributes, and the relationships between node attributes, hypotheses about the influencing factors of knowledge diffusion that need to be empirically analyzed are proposed. The third part of the framework is gradually constructing ERGMs based on the hypotheses on the seconds parts, and applying the statnet library in *R* language to solve the constructed ERGMs, and empirically analyzing the various influencing factors of knowledge diffusion.

3.2 Construction of Citation Networks

Citation networks generally utilize papers as nodes and construct a network via citation relationships, which can be viewed as the dissermination of knowledge from one author to another. The primary entity of knowledge diffusion is the author of the paper; therefore, we consider the first author of a paper as a node in the knowledge diffusion network, and the citation relationship between the published papers of the first author as the directed edge of the knowledge diffusion network. The edge indicates that the knowledge has diffused from the first author of the cited document to the first author of the citing document.

3.3 Knowledge Role Identification

Different behaviors endow knowledge nodes with different knowledge roles. The main behavior of knowledge seekers is to seek help and conduct inquiries, while the main behavior of knowledge contributors is to provide help and share knowledge. Brokers are intermediaries who facilitate the transfer of knowledge between different individuals, groups, or organizations. They often connect seekers with contributors and help bridge gaps between disparate knowledge domains. Lurkers are individuals who observe and consume knowledge without actively participating in discussions or contributing content.

In the citation network, by calculating the in degree and out degree of each node, we can define the four roles mentioned above (as shown in Table 1).

Table 1 Definition of knowledge roles

Description	Role
In-degree greater than or equal to 1, out-degree is 0	contributor
In-degree is 0, out-degree is 0	lurker
In-degree greater than or equal to 1, out-degree greater than or equal to 1	broker
In-degree is 0, out-degree greater than or equal to 1	seeker

3.4 Assumptions About Impact Factors

a) Hypothesis on endogenous structure

The endogenous structural effect of a network arises from its self-organizing process. Peng[29] introduced geometrically weighted dyadwise shared partners (GWDSP) and geometrically weighted edgewise shared partners (GWESP) into the ERGM, and identified the transfer ternary as an important force in facilitating the formation of citation links between journals. An and Ding[30] constructed a citation network comprising the top authors in the field by integrating the GWESP and GWDEP into ERGM to account for the transitivity, and incorporating geometrically weighted in-degree distributions (GWIDEGREE) and geometrically weighted out-degree (GWODEGREE) distributions. Their model fitting resulys indicate that the endogenous network formation process plays an important role in the formation of citations.

We select the endogenous structural effect as a factor influencing knowledge diffusion to study its impact on the formation and evolution of diffusion relationships in the knowledge diffusion network. We have the following hypothesis:

H1 The endogenous structural effects of networks have a significant impact on the formation of diffusion relationships in citation networks.

The configurations for directed network reflect endogenous structural variables including transitivity, activity, popularity, and connectivity[31]. Therefore, these 4 types of effects are selected as endogenous structural variables, with the network edges as the control variables to build the main effect model. The statistical measures of these 5 variables are shown in the table 2.

In citation networks, transitive closure is characterised by two aspects. First, transitive closure is an arc added to the 2-path, making missing links visible which enhancing the internal relationships of the structure, and this feature can be used to analyse the development path of the technology[32]. Second, the node degree distribution in the transfer closure construction is not uniform, with certain nodes have in-degree advantages, exhibiting better performance in knowledge flows than popularity, and the transfer effect can be used to identify the source of knowledge flow processes[33,34].

In network analysis, degree distribution effects refer to the influence of sub-structural features of different degrees on the formation of network relationships, assuming that other factors in the network are held constant[35]. For example, in a citation network, some documents are highly cited, whereas others are not. This variation in the citation degree can be observed in the disparity of the centrality distribution[35]. Because citation networks are directed, both in-degree centrality (aggregation) and out-degree centrality (expansion) must be taken into account when constructing citation-related networks.

The connectivity effect refers to a network in which two nodes are connected through one or more shared nodes so that they can efficiently and accurately find potential collaborators[36]. Connectivity refers to the GWESP distribution, and the effect assumes that the emergence of a 2-path structure influences on the formation of citation relationships, while holding all other factors in the network constant. The 2-path structure, frequently encountered in citation network research, is similar to a simplified indirect citation structure. This structure is characterized by two papers form a connected 2-path structure through a third paper, and there is no direct citation relationship between the two papers. Thus, the 2-path structure helps to find potentially similar literature.

Circularity is the addition of directly connected dependencies based on binary partners (popularity, transitivity, and activity), forming a transitive closed triplet[37]. This structure is termed a shared partner distribution because it facilitates the formation of closed circular relationships that enhance network connectivity.

Therefore, this study refines hypothesis H1 into the following hypotheses:

H1a There is a correlation between the degree distribution of the network and the likelihood of citation relationship formation in the citation network, and the popularity and activity of the network will positively affect knowledge diffusion in the network.

H1b There is a correlation between the transferability of the network and the possibility of forming citation relationships in the citation network, which will have a positive impact on the knowledge diffusion in the network.

H1c There is a correlation between the connectivity of the network and the possibility of forming citation relationships in the citation network, which will have a positive impact on the knowledge diffusion in the network.

b) Hypothesis on knowledge roles

According to our definition in section 3.3, lurkers are essentially independent of other nodes and have no association with them. Therefore, our empirical research does not consider the role of lurkers. As discussed in section 2, the distribution of the three categories of knowledge roles - contributor, seeker, and broker - will have an impact on the spread of knowledge[21,38]. However, we want to analyzing the following issues: 1) Do these three different knowledge roles have a certain distribution that can promote and strengthen the process of knowledge diffusion? 2) If so, what is the optimal distribution for knowledge diffusion? Therefore, we propose the following hypothesise:

H2 The distribution of different knowledge roles affects the extent of knowledge diffusion in citation networks.

Further we decompose hypothesis H2 into the following hypotheses by knowledge roles:

H2a There is a relationship between the number of contributors and the likelihood of citation relationships forming in citation networks.

H2b There is a relationship between the number of seekers and the likelihood of citation network citation relationship formation.

H2c There is a relationship between the number of brokers and the likelihood of citation network citation relationship formation.

H2d There is a relationship between the number of lurkers and the likelihood of citation network citation relationship formation.

c) Hypothesis on assortativity of network nodes

Assortativity is the phenomenon in which nodes in a network tend to form connections with other nodes that are similar to themselves, also known as assortative mixing. Newman[39] posited that network assortativity refers to the tendency for nodes in a network with multiple connections to connect with other nodes that also have multiple connections. Brede and Newth[40] proposed a method for calculating the assortativity of nodes in a network using their degree values. The basic assumption is that nodes with higher degree values are more likely to connect with other nodes with higher degree values. Noldus and Mieghem[41] investigated the concept of assortativity in complex networks, describing how nodes tend to connect with other nodes that are similar or different based on certain characteristics, such as degree. Zhang et al.[42] investigated the impact of two network mechanisms - transmission mechanisms and preferred linking mechanisms - alongside author attributes such as author productivity, influence, research topic, and gender, on the formation of author collaborative relationships.

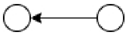
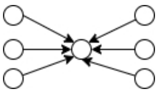
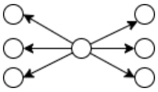
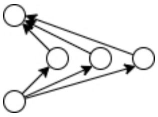
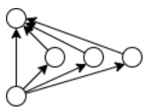
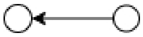
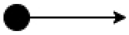
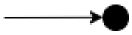
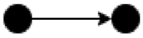
In summary, the assortativity of nodes leads them to preferentially link with nodes exhibiting similar attributes, as represented in the network's interaction effect. Therefore, the interaction effect between nodes is selected as one of the factors influencing knowledge diffusion in the citation network. Based on this, the following hypothes is proposed:

H3 The homogeneity between knowledge roles can affect the level of knowledge diffusion in citation networks.

d) Description of variables

The variables in the above hypotheses and their descriptions are shown in Table 2.

Table 2 Variables in hypotheses

	Variable	Structural diagram	Expression	Description
Statistics item	edges		$\sum_{i,j} x_{i,j}$	Number of edges in the citation network
	popularity		$\sum_{j=0}^{n-1} \mathrm{e}^{-aj} d_j^i n$	Is the network centered on in-degree penetration and do nodes have similar levels of popularity
	activity		$\sum_{j=0}^{n-1} \mathrm{e}^{-aj} d_j^o ut$	Is the network centered on out-degree penetration and do nodes have similar levels of activity
	transitivity		$z_T(x,\lambda) = 3T_1(x) - \frac{T_2(x)}{\lambda} + \cdots + (-1)^{n-3}\frac{T_{n-2}(x)}{\lambda^{n-3}}$	Whether or not there is a large number of passing triangular clusters in the network
	connectivity		$z_P(x,\lambda) = P_1(x) - \frac{P_2(x)}{\lambda^2} + \cdots + (-1)^{n-3}\frac{P_{n-2}(x)}{\lambda^{n-2}}$	Describe the trend of passing referential relationships through multiple paths
Node attributes	main effect		$\sum_{i,j} x_{ij}(y_i + y_j)$	The effect of node attributes on the formation of citation network relationships
	out-degree		$\sum_{ij} y_{ij}\delta_i$	Nodes with certain attributes are more likely to be cited
	in-degree		$\sum_{ij} y_{ij}\delta_j$	Nodes with certain attributes are easier to induce
Inter-node interaction effects	homogeneous		$\sum_{ij} x_{ij} y_i y_j$	Two nodes with the same attributes are more likely to form a reference relationship

3.5 Exponential Random Graph Model of Knowledge Diffusion

Next, as illustrated in Figure 1, we will build four empirical models based on ERGM: Null model, network structure model, node main effect model, and interaction effect model. These models will be used to investigate the impact of the endogenous structure effect, node main effect, and node interaction effect in citation networks.

The null model is a basic random graph model comprised only of the edges of a single statistical term network. While not considering complex dependency assumptions, it serves as a benchmark for evaluating the goodness-of-fit when constructing complex models, and describes the density characteristics of the observed network.

The network structure model builds on the null model by including the network endogenous structure effect variable to investigate the influence of the network's self-organisation process on the formation of network relationships, and "stucture" refers to the statistical term of the network endogenous structure effect.

The main effect model adds node attribute variables to the network structure model to examine the impact of node knowledge role attributes on the formation of citation relationships in the network, and we use "nodefactor" to denote the statistical terms of subtypes of variables in the network. The node attributes in this paper denote the knowledge roles of nodes in the citation network.

The interaction effect model expands on the main effect model by including the interaction terms, primarily focused on the impact of the homogeneity or heterogeneity in the attribute features of two nodes on the formation of network relationships. We use "absdiff" to denote the heterogeneity variable in the network of statistical terms and "nodematch" to denote the heterogeneity variable in the network of statistical terms. The construction process of the model is shown in Table 3.

Table 3 ERGMs building

	Null model	Network architecture model	Main effect model	Interaction effect model
ERGMs	ergm ∼ edges	ergm ∼ edges + structure	ergm ∼ edges + structure + nodefactor	ergm ∼ edges + stucture + nodefactor + absdiffcat+ nodematch
Factors/ variables	Number of edges	1. Number of edges 2. Network endogenous structural variables	1. Number of edges 2. Network endogenous structural variables 3. Node's knowdege roles	1. Number of edges 2. Network endogenous structural variables 3. Node's knowdege roles 4. Homogeneity of knowledge roles

4 Results and Analysis

4.1 Data Collection and Pre-Processing

With the rapid development of big data technology, knowledge generation and dissemination within the field have become increasingly important[43]. This study uses research data in the field of “Big Data” as an example to study the impact of knowledge role on knowledge diffusion in citation networks.

The literature data for this study is sourced from the Web of Science database. We employ TS = “big data” OR TS = “mega data” OR TS = “bigdata” OR TS = “megadata” as the subject for query search, and selecting four types of literature: Article, proceeding paper, review, and book review, resulting in 80 source data files in text format (each containing 500 records), which are metadata covering the period from 2008 to 2019. These metadata encompass details such as the DOI, title, author, institution, publication date, publication journal, and citation relationship. The aforementioned metadata is parsed using the Python programming language, resulting in a dictionary relationship table formatted as {keyword: attribute value}.

The solution of ERGM is very computationally intensive. Without better solving tools, for ordinary computer workstations, the StatNet library in *R* language can only solve networks with about 1000 nodes. As the network size expands, the computational complexity and the number of network nodes increase exponentially. Therefore, we need to limit the network size so that the model can be solved. We filter nodes by setting a threshold for node degree, thus obtaining a core network.

4.2 Data Descriptive Statistics

The core citation network data in the field of big data from 2008 to 2019 was selected for ERGM analysis using the statnet package in *R*. The 2008–2019 Citation Networks in the Field of Big Data consists of 1513 nodes, 1897 edges, and a network density of 0.000829. Using the knowledge role identification method in the citation network as defined in the previous section, the roles are discerned, revealing a total of 814 contributors, 205 brokers, and 494 seekers in the network.

Figure 2 illustrates a network graph that depicts the features of knowledge roles, with white

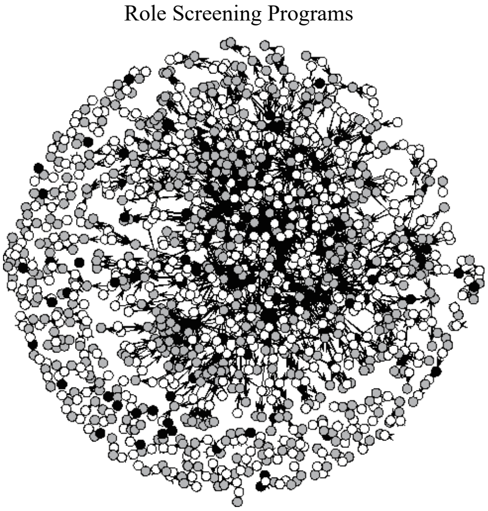


Figure 2 Knowledge role characterisation network

nodes representing contributors, grey nodes representing seekers, and black nodes representing brokers. The graph shows no obvious aggregation phenomenon, and most of the connecting edges between nodes occur between nodes of different categories. This network graph implies that the probability of generating knowledge diffusion between different knowledge roles in the citation network exceeds the probability of randomly establishing relationships between knowledge roles.

A mixed matrix can be used to analyze possible combinations between different levels of categorical attribute variables. Table 4 shows the mixed matrix of node knowledge roles. There are 517 binary pairs between contributors and brokers, and 904 binary pairs between brokers and searchers. The number of binary pairs between contributors, brokers, and seekers is 517 and 904, respectively. The number of binary pairs between brokers, brokers, and seekers is 172 and 304, respectively. Nodes with different knowledge roles are more likely to form connections, indicating differentiated connection patterns.

Table 4 Mixed matrix of node knowledge role

	Contributor	Broker	Seeker
Contributor	0	517	904
Broker	0	172	304

4.3 Model Construction and Result Analysis

We construct a null model based on the aforementioned dataset and subsequently incorporate network structure effects, node main effects, and interaction effects as explanatory variables into the ERGM to analyze the formation of citation relationships and their influencing factors in the citation network of the big data field. Meanwhile, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are adopted to assess the model's goodness of fit; lower the AIC and BIC values indicate a better model fit to the data, suggesting that the model aligns more closely with the observed network. The results of the models estimated in this paper are shown comprehensively in Table 5. The findings are elaborated further in the remainder of this section.

An ERGM is a probability distribution with the form:

$$p(X = x \mid \theta) = \frac{1}{N} \exp\{\theta_1 c_1(x) + \theta_2 c_2(x) + \cdots + \theta_p c_p(x)\}. \tag{1}$$

The null model estimation results demonstrate that the coefficient of the statistical term of the number of edges is negative (-7.094) and significant. This suggests that the citation network has a comparatively low number of edges compared to a random network, and the network has a density of 50% or less, aligning with the typical characteristics of actual observation networks.

The network structure model includes GWDSP, GWESP, GWIDEGREE and GWODEGREE. In a directed weighted network, GWDSPs are used to measure the probability that nodes without direct links are interconnected via indirect connections. The GWDSP statistical term is negatively significant, indicating that links that create open triangles in knowledge diffusion networks are less likely to be formed in the network, and there are challenges in forming

Table 5 ERGM model estimation results

Variable	Null model	Network structure model	Main effect model	Interaction effect model
edges	−7.094***	−5.843***	−5.860***	−5.467***
Network structure effects				
GWDSP		−0.175***	−0.150***	−0.225***
GWESP		2.990***	1.202***	1.314***
GWIDEGREE		−2.525***	−1.625***	−1.296***
GWODEGREE		−0.345***	−0.046***	−0.007***
Main effects				
nodeicov.indegree			0.066***	0.070***
nodeocov.outdegree			0.098***	0.094***
Nodefacto.role-Contributor			−0.550***	−0.059***
Nodefacto.role-Seeker			−0.403***	−0.473***
Interaction effects				
Nodematch.role				−1.557***
Goodness of fit				
AIC	23152	21747	20104	19159
BIC	23164	21809	20214	19282

direct connections between two nodes lacking a prior citation relationship through a common node. This suggests a limitation on knowledge diffusion within citation networks; specifically, the probability of creating a citation relationship between two authors is constrained by the number of nodes they share neighbours with. Therefore, authors with fewer shared neighbours may face more limitations, potentially leading to fewer connections and impacting the knowledge diffusion process, thus supporting hypothesis H1c.

The effect of GWESP is positively significant in the citation network, indicating that it is likely to form transmission triangles. This suggests that authors who have previously made a direct citation (cited) are also likely to form a new citation relationship with each other through the common citing (cited) authors, indicating that there is a more intensive and extensive transfer and exchange of knowledge and information in the network. In this case, knowledge diffusion in the network will have a greater depth, breadth and involvement. Hypothesis H1b is supported.

GWIDEGREE and GWODEGREE can be used to denote popularity and activity, respectively, and serve as measures of the central tendency of the citation distribution in a citation network. The estimates of the structural model for popularity and activity are both significantly negative. Negative popularity indicates that the probability of establishing a citation relationship between a pair of nodes is lower than the probability of a randomly occurring citation relationship, and there are no core authors in the sample network in the area of big data. Activity reflects the central tendency of network scalability, while negative activity indicates that literature seekers in the sample network are not active in the citation network. The GWIDEGREE coefficient value is -2.525^{***}, indicating that the high in-degree nodes tend to have lower geometric weights, meaning that the quality and influence of the cited articles (nodes) obtained are relatively low, and correspondingly, the influence and knowledge influence of the nodes are also low. The GWODEGREE coefficient value of -0.345^{***} indicates that the nodes with high out-degree exhibit lower geometric weights, which means that the quality and influence of the cited articles obtained by the node are relatively low, and correspondingly, the overall influence and knowledge diffusion ability of the node on the network are also relatively low.

Regarding knowledge diffusion in the citation network, higher in-degree nodes exhibit lower the quality, influence and knowledge diffusion capabilities, which may therefore restrict the outward transmission of knowledge from these nodes. Correspondingly, the quality and influence of articles cited by high in-degree nodes are also reduced, so the information obtained from articles cited by these nodes may also be somewhat limited. However, it should be noted that such effects are relative, that is, compared to nodes with high geometric weights in the network, these nodes have less influence but still have a certain role and influence in the whole network, supporting hypothesis H1a.

The results of the main effects model show that the indegree coefficient is positively significant at 0.066, indicating that the number of references (in degree) of each node in the network increases by one, and the probability of being connected to other nodes increases by about 6.6%. This suggests that more highly cited nodes are more likely to be connected to other nodes. This is consistent with the fact that in academia, scholars with higher citation counts are more likely to collaborate or engage with other scholars due to their more prominent contributions and established reputations. In the network, nodes with higher citation counts have possess an increased likelihood of connecting with other nodes, thereby enhancing their opportunities to spread their ideas or research findings to other nodes. This facilitates knowledge diffusion and exchange, strengthens the connection between nodes, and consequently promoes the development of the corresponding academic field. The out-degree coefficient is 0.098, which is positively significant, indicating that the number of references (out-degree) for each node in the network increases by one, the probability of connecting to other nodes increases by about 9.8%. Higher outdegree coefficients indicate that a node is more active in issuing citations and that these citations, in turn, amplify its influence on other nodes, leading to a more significant impact in the knowledge dissemination process. Although the coefficient values for both indegree and outdegree are relatively small (0.66 and 0.098), their significance suggests that we need to consider the indegree and outdegree attributes of the nodes in order to more accurately predict the relationships between nodes in the network.

In terms of the statistics of the node knowledge roles, the broker served as the reference group and the coefficient of the seeker node was negative, indicating that the seeker nodes have less influence in knowledge dissemination compared to the broker. Seekers typically search for original knowledge sources and acquire and disseminate knowledge in the process, but they may lack sufficient connections in the network. This means that seekers may lack sufficient opportunities to pass on the knowledge they find to other nodes compared to brokers. The coefficient of contributor nodes is significantly negative compared to broker nodes, which implies that contributors exert less influence than brokers in the diffusion of knowledge in the network. Furthermore contributor nodes may not be sufficiently active or influential to effectively facilitate the knowledge flow and exchange. This result may indicate that contributor nodes have limitations in terms of connectivity and authority in the network. Therefore, since contributor nodes play an important role in the dissemination, acceptance, and application of knowledge in the network, the result (the coefficient of contributor nodes compared to broker nodes is significantly negative) indicates that knowledge diffusion within the network is constrained. Based on the analysis of the above results, this may indicate that the role of contributors and seekers is not as important as that of brokers in citation networks. Therefore, there is a greater need to leverage brokers to facilitate knowledge diffusion, supporting hypotheses H2a, H2b, and H2c.

According to the interaction effect model, the homogeneity effect of node knowledge role attributes is significantly negative. This suggests that authors in the sample citation network are more likely to establish citation relationships with authors whose node knowledge roles differ from their own. This finding suggests that the diffusion of knowledge in the network is facilitated by communication between various knowledge roles. This finding does not support the hypothesis H3.

Based on the above analysis, the topology of the citation network may have a significant impact on the dissemination and diffusion of knowledge. For example, strong connections in the network may facilitate knowledge and information diffusion, while isolated nodes in the network may impede the diffusion of knowledge. The knowledge roles of nodes may also have an impact on the formation of citation relationships in citation networks in the area of big data in 2008–2015. An increase in the number of contributors and seekers in a citation network may facilitate the dissemination of knowledge. However, the quality of the authors' papers is also important, and this attribute is independent of quantity. Contributors and seekers must be able to provide valuable knowledge and facilitate its dissemination.

4.4 Model Fitting and Model Diagnostics

This paper evaluates the model fitting results using the AIC and BIC. In the model constructed in Section 4.3, the AIC value of the null model is 23152, the BIC value is 23164. The AIC value of the interaction effect model is 19159, the BIC value is 19282. The AIC and BIC values of the interaction effect show a substantial improvement compared to the null model. This indicates that the statistical terms in the interaction effect model play an important role in improving the goodness of fit of the ERGM, and that the interaction effect model provides the optimal fit.

Model diagnostics can assist in determining whether the estimation algorithm has converged whether an approximate degradation problem exists, thereby indicating whether it is the model itself or the model evaluation setup conditions that need to be adjusted. Graphical display of

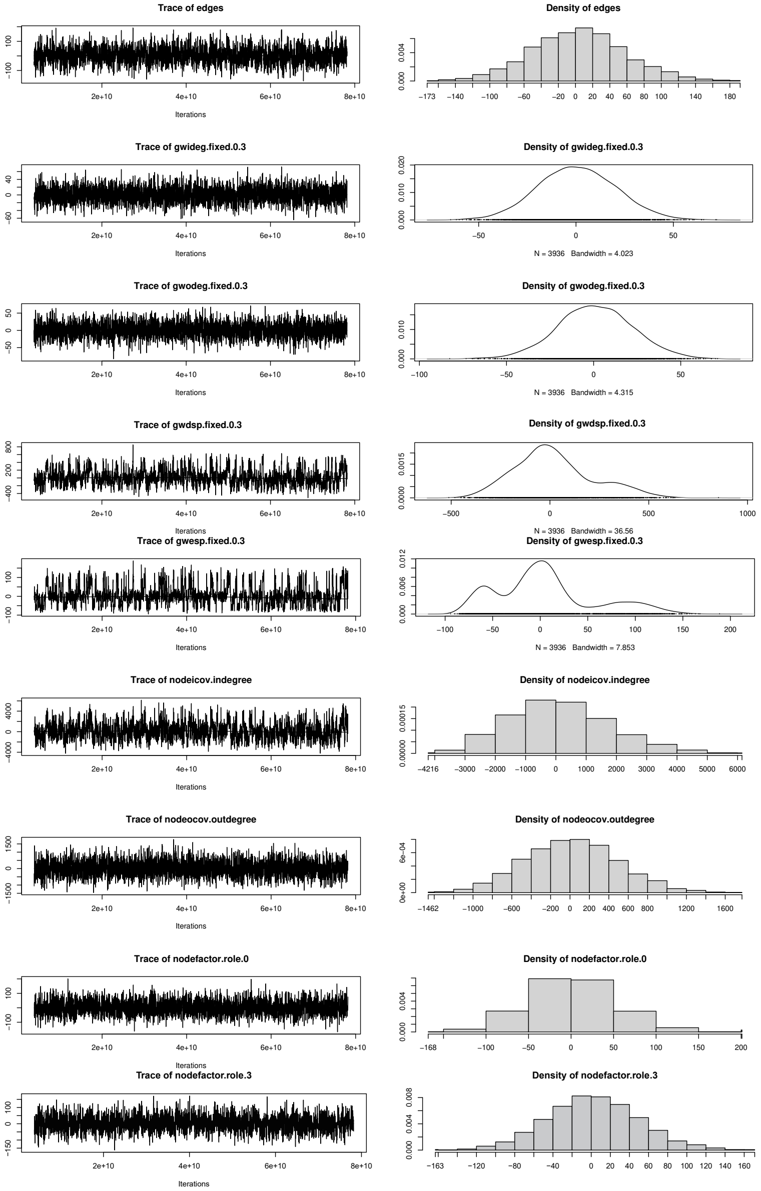


Figure 3 Interaction model graphical diagnostic results (partial)

Monte Carlo Markov Chain (MCMC) diagnostics is an effective test method. Figure 3 shows the iteration of the interaction effect model in the final stage. The statistical chart on the left side of Figure 3 takes each statistical item in the model as a unit, and utilizes the MCMC chain to build a time series to display the change of statistical items. The statistical chart on the right side shows the histogram of the corresponding MCMC chain. If the model is able to converge, the graph of each statistical term in the model will show random variation around 0, where 0 represents the statistical value of the corresponding statistical term of the observation network. In the aforementioned model, each statistical term exhibits random variation around 0. Overall, the results of the model diagnostics show that the model is stable.

5 Discussion

Our empirical results support Hypothesis 1, which states that the endogenous structural effects of the network have a significant impact on the formation of diffusion relationships in citation networks, and Hypotheses 2a, 2b, and 2c, which state that there is a correlation between the number of contributors, seekers, and brokers and the likelihood of citation relationship formation in the citation network. The empirical results did not support Hypotheses 3, i.e., there was no evidence that homogeneity among knowledge roles has an effect on knowledge diffusion.

Yang et al. simulated the effect of different network structures on knowledge diffusion[44]. Our empirical results for Hypothesis 1 support Yang et al.'s simulation results in this regard. Wu et al.[45] found that posters and lurkers exhibit different behaviors in terms of the motivating factors for knowledge sharing in an online community. This conclusion precisely supports our empirical results for hypothesis H2, indicating that different knowledge roles have varying impacts on the formation of networks.

In terms of homogeneity, our empirical results do not support Hypothesis 3. In terms of homogeneity, our empirical results do not support Hypothesis 3. Since there are no studies similar to our empirical analysis of knowledge diffusion in terms of knowledge roles, we refer to empirical analyses in other fields, for example, interdisciplinary knowledge diffusion with disciplinary dependence is to show that nodes with the same disciplinary attributes cannot promote knowledge diffusion[42], similar to our conclusion that we do not support the assumption of homogeneity of node roles.

5.1 Theoretical and Practical Implications

This paper proposes a research framework for empirically investigating the impact of network structure and knowledge roles on knowledge diffusion, and establishes ERGMs based on citation networks to investigate the relative magnitude of the influence of knowledge roles, specifically knowledge seekers, brokers, and contributors, on the knowledge diffusion. This approach offers a novel perspective on the study of influencing factors in the field of knowledge diffusion.

Our research enhances the comprehension of how various knowledge roles affect knowledge transmission and may impact scholars' emphasis on these roles. This study's results can offer scholars novel insights into the influence of knowledge roles on knowledge spread. The findings in Table 5 regarding the impact of node roles indicate that brokers exert a more significant influence on knowledge diffusion than other roles, underscoring the importance of this discovery for knowledge managers in establishing suitable guidelines to maintain broker engagement.

The heterogeneity of the degree to which knowledge roles impact the process of knowledge diffusion is another main topic of the research. The heterogeneity of the node attributes implies that it is emphasizing to those responsible for knowledge management that knowledge diffusion is more likely to occur between various knowledge roles. In practical application, knowledge managers can focus on selecting different knowledge roles as diffusion subjects at different stages of network development, so as to promote the diffusion of knowledge in the network.

5.2 Limitations

Our research experience exhibits that using ERGM to study networks presents significant challenges. As is well known, due to its reliance on the MCMC for estimation, ERGM can be computationally slow and less adaptable to large networks. The ergm package inside the R language's statnet framework can only accommodate networks with about 1000 nodes. For larger networks, the estimation process is very slow, often requiring weeks, and prone to system crashes that may not yield valid results. The hardware utilized in this research includes a Lenovo ThinkStation P520c with 48G RAM, an Intel Xeon(R) W-2133 CPU, and a Quadro P2200 graphics card manufactured by NVIDIA Corporation. The performance of the devices is adequate for the common computational tasks required. However, when our citation network exceeds 1500 nodes, obtaining results becomes challenging. This is because ERGM relies on MCMC to simulate the network for estimation, and the computational load grows exponentially with the number of nodes in the network. Therefore, this paper is limited to constraining the network to an acceptable range based on the data. We have initiated efforts to develop our own modelling algorithm for ERGM using parallel computing, GPU computing, and various techniques to develop algorithmic tools that will facilitate larger-scale network computation.

In addition, the empirical ERGM in this paper only considers the knowledge roles of the nodes in terms of their attributes, neglecting other contributing factors such as geographic proximity, disciplinary matching, institutional matching, and position matching. Future research should focus on refining the selection of influencing factors in knowledge diffusion networks and exploring how to apply exponential random graphs in large-scale networks.

6 Conclusion

This paper develops the ERGMs in terms of the endogenous structure of the network, the attributes of the network nodes, and the homogeneity of the nodes, and investigates the relative magnitude of the impact of different knowledge roles on knowledge diffusion. Our results indicate that: 1) In the sample network, links forming open triangles are less prevalent, and the probability of forming a citation relationship between two authors is limited by the number of nodes in their shared neighbourhood. As a result, authors with fewer shared neighbors may encounter greater limitations and fewer connections, thereby affecting the process of knowledge diffusion. 2) Citation links that transmit triangles are likely to form in the network, which suggests that new citation relationships are also likely to form between authors who have previously made a direct citation (cited) through a common citing (cited) author. This may indicate that transferring triangular configurations can facilitate knowledge diffusion. 3) The estimates of both popularity and activity in the structural model are negatively significant, there are no core authors in the sample network of the field of big data, and literature seekers in the sample network do not behave actively in the citation network. 4) Seeker and contributor nodes have less influence on knowledge diffusion than brokers. Seekers may not have enough opportunities to share the knowledge they find with other nodes. The contributor nodes are not active or influential enough to effectively promote the flow and exchange of knowledge. This result may indicate the limitations of contributor nodes in terms of connectivity and authority in the network. Since contributor nodes play an important role in the dissemination, this may impose some limitations on the diffusion of knowledge in the network, where more reliance on brokers is needed to facilitate knowledge diffusion. Finally, the effectiveness of the model was confirmed through model fitting and model diagnosis.

References

[1] Hassan S U, Safder I, Akram A, et al. A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis. Scientometrics, 2018, 116(2): 973–996.

[2] Jiang S, Chen H. Examining patterns of scientific knowledge diffusion based on knowledge cyber infrastructure: A multi-dimensional network approach. Scientometrics, 2019, 121(3): 1599–1617.

[3] Jong J Y, Wu W W, So S R. A new model for competitive knowledge diffusion in organization based on the statistical thermodynamics. Advances in Mathematical Physics, 2020, 2020(1): 8491516.

[4] Mu J, Tang F, MacLachlan D L. Absorptive and disseminative capacity: Knowledge transfer in intra-organization networks. Expert Systems with Applications, 2010, 37(1): 31–38.

[5] Liu X, Jiang S, Chen H, et al. Modeling knowledge diffusion in scientific innovation networks: An institutional comparison between China and US with illustration for nanotechnology. Scientometrics, 2015, 105(3): 1953–1984.

[6] Kim H, Park Y. Structural effects of R&D collaboration network on knowledge diffusion performance. Expert Systems with Applications, 2009, 36(5): 8986–8992.

[7] Tang F, Mu J, Maclachlan D. Implication of network size and structure on organizations' knowledge transfer. Expert Systems with Applications, 2008, 34(2): 1109–1114.

[8] Havakhor T, Soror A A, Sabherwal R. Diffusion of knowledge in social media networks: Effects of reputation mechanisms and distribution of knowledge roles. Information Systems Journal, 2018, 28(1): 104–141.

[9] Quigley N R, Tesluk P E, Locke E A, et al. A multilevel investigation of the motivational mechanisms underlying knowledge sharing and performance. Organization Science, 2007, 18(1): 71–88.

[10] Bielak A T, Campbell A, Pope S, et al. From science communication to knowledge brokering: The shift from 'science push' to 'policy pull'. Communicating Science in Social Contexts: New Models, New Practices, 2008: 201–226.

[11] Nam K K, Ackerman M S, Adamic L A. Questions in, knowledge in? A study of Naver's question answering community. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2009: 779–788.

[12] Preece J, Nonnecke B, Andrews D. The top five reasons for lurking: Improving community experiences for everyone. Computers in Human Behavior, 2004, 20(2): 201–223.

[13] Turnhout E, Stuiver M, Klostermann J, et al. New roles of science in society: Different repertoires of knowledge brokering. Science and Public Policy, 2013, 40(3): 354–365.

[14] Zhang J, Ackerman M S, Adamic L. Expertise networks in online communities: Structure and algorithms. Proceedings of the 16th International Conference on World Wide Web, 2007: 221–230.

[15] Qiao T, Shan W, Zhang M, et al. How to facilitate knowledge diffusion in complex networks: The roles of network structure, knowledge role distribution and selection rule. International Journal of Information Management, 2019, 47: 152–167.

[16] Chen C, Hicks D. Tracing knowledge diffusion. Scientometrics, 2004, 59(2): 199–211.

[17] Liu X, Li Y. A network-based analysis of the knowledge diffusion system in science. Scientometrics, 2016, 108(1): 365–386.

[18] Cowan R, Jonard N. Network structure and the diffusion of knowledge. Journal of Economic Dynamics and Control, 2004, 28(8): 1557–1575.

[19] Singh J. Collaborative networks as determinants of knowledge diffusion patterns. Management Science, 2005, 51(5): 756–770.

[20] Abramo G, D'Angelo C A, Di Costa F. The role of geographical proximity in knowledge diffusion, measured by citations to scientific literature. Journal of Informetrics, 2020, 14(1): 101010.

[21] Welser H T, Gleave E, Fisher D, et al. Visualizing the signatures of social roles in online discussion groups. Journal of Social Structure, 2007, 8(2): 1–32.

[22] Marett K, Joshi K D. The decision to share information and rumors: Examining the role of motivation in an online discussion forum. Communication of the Association for Information Systems, 2009, 24(1): 4.

[23] Nonnecke B, Preece J. Lurker demographics: Counting the silent. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2000: 73–80.

[24] Gray B. Informal learning in an online community of practice. Journal of Distance Education, 2004, 19(1): 20–35.

[25] Ridings C, Gefen D, Arinze B. Psychological barriers: Lurker and poster motivation and behavior in online communities. Communication of the Association for Information Systems, 2006, 18(1): 16.

[26] Cassi L, Corrocher N, Malerba F, et al. The impact of eu-funded research networks on knowledge diffusion at the regional level. Res. Eval., 2008, 17(4): 283–293.

[27] Macy M W, Willer R. From factors to actors: Computational sociology and agent-based modeling. Annu. Rev. Sociol., 2002, 28(1): 143–166.

[28] Wang J, Zhang L. Proximal advantage in knowledge diffusion: The time dimension. Journal of Informetrics, 2018, 12(3): 858–867.

[29] Peng T Q. Assortative mixing, preferential attachment, and triadic closure: A longitudinal study of tie-generative mechanisms in journal citation networks. Journal of Informetrics, 2015, 9(2): 250–262.

[30] An W, Ding Y. The landscape of causal inference: Perspective from citation network analysis. The American Statistician, 2018, 72(3): 265–277.

[31] Hunter D R, Krivitsky P N, Schweinberger M. Computational statistical methods for social network models. Journal of Computational and Graphical Statistics, 2012, 21(4): 856–882.

[32] Wang J C, Chiang C H, Lin S W. Network structure of innovation: Can brokerage or closure predict patent quality? 2009 42nd Hawaii International Conference on System Sciences. IEEE, 2009: 1–10.

[33] Batagelj V. Efficient algorithms for citation network analysis. arXiv preprint cs/0309023, 2003.

[34] Hung S W, Wang A P. Examining the small world phenomenon in the patent citation network:

A case study of the radio frequency identification (RFID) network. Scientometrics, 2010, 82(1): 121–134.

[35] Leskovec J, Kleinberg J, Faloutsos C. Graphs over time: Densification laws, shrinking diameters and possible explanations. Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, 2005: 177–187.

[36] Morris M, Handcock M S, Hunter D R. Specification of exponential-family random graph models: Terms and computational aspects. Journal of Statistical Software, 2008, 24(4): 1–24.

[37] Cranmer S J, Desmarais B A, Menninga E J. Complex dependencies in the alliance network. Conflict Management and Peace Science, 2012, 29(3): 279–313.

[38] Caimo A, Lomi A. Knowledge sharing in organizations: A Bayesian analysis of the role of reciprocity and formal structure. Journal of Management, 2014, 40(6): 1587–1609.

[39] Newman M E J. Assortative mixing in networks. Physical Review Letters, 2002, 89(20): 208701.

[40] Brede M, Newth D. Patterns in syntactic dependency networks from authored and randomised texts. CS2004. CUQ Press, 2004: 1–17.

[41] Noldus R, Mieghem V. Assortativity in complex networks. Journal of Complex Networks, 2015, 3(4): 507–542.

[42] Zhang C, Bu Y, Ding Y, et al. Understanding scientific collaboration: Homophily, transitivity, and preferential attachment. Asso for Info Science & Tech, 2018, 69(1): 72–86.

[43] Liu G Y. Understanding digital platform standardization and sustainability within big data context: Case of a platform business in China. Aachen: Annual Conference of European Academy for Standardization (EURAS), 2020.

[44] Yang G Y, Hu Z L, Liu J G. Knowledge diffusion in the collaboration hypernetwork. Physica A: Statistical Mechanics and Its Applications, 2015, 419: 429–436.

[45] Wu C, Hill C, Yan E. Disciplinary knowledge diffusion in business research. Journal of Informetrics, 2017, 11(2): 655–668.