Formation Mechanisms of Patent Citation Networks Based on Exponential Random Graph Models: A Case Study of Nelarabine Drug (Postprint)

Authors: Yang Guancan, Zhanlin Liu, Li Gang

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] The formation of patent citation relationships constitutes a crucial issue for understanding innovation networks. Traditional regression models, which assume independence among observations, are unable to incorporate structural effects of networks into the model to provide comprehensive statistical inference. The Exponential Random Graph Model (ERGM) represents an innovative statistical inference approach that integrates three types of features: attribute characteristics, self-organization characteristics, and network synergy characteristics. [Method/Process] This study employs the patent citation network of nelarabine drugs as the research object and utilizes ERGM to systematically examine five mechanisms influencing patent citation relationships: main effects of patent attributes; time difference effects of patent citations; closure effects of patent citation relationships; transitivity effects of patent citation relationships; and network synergy effects of patent citation relationships. [Results/Conclusion] All five mechanisms contribute to the formation process of patent citation relationships for nelarabine drugs. However, three effects demonstrate the most pronounced influence: synergy effects of shared inventor relationships, synergy effects of shared patent family relationships, and transitivity effects. Additionally, certain auxiliary mechanisms also impact the formation of patent citation relationships, such as citation lag, number of claims, and number of references.

Full Text

A Study on the Formation Mechanisms of Patent Citation Relationships Based on Exponential Random Graph Models: A Case Study of the Nelarabine Drug

Yang Guancan[1], Liu Zhanlin[2], Li Gang[3]

[1]School of Information Resource Management, Renmin University of China, Beijing 100872
[2]Department of Industrial and Systems Engineering, University of Washington, Seattle 98105
[3]School of Information Management, Wuhan University, Wuhan 430072

Abstract

[**Purpose/Significance**] The formation of patent citation relationships is a critical issue for understanding innovation networks. Traditional regression models rely on independence assumptions that cannot integrate network structural effect factors into the model to provide comprehensive statistical inference. The Exponential Random Graph Model (ERGM) represents an innovative statistical inference methodology capable of simultaneously examining three types of features: attribute characteristics, self-organizational features, and network covariate features. [**Method/Process**] This study takes the patent citation network of the Nelarabine drug as its research object and employs ERGM to systematically test five mechanisms influencing patent citation relationship formation: main effects of patent attributes; difference effects of patent citation timing; convergence effects of patent citation relationships; transitivity effects of patent citation relationships; and network covariate effects of patent citation relationships. [**Result/Conclusion**] All five mechanisms play roles in the formation process of patent citation relationships for the Nelarabine drug. However, three effects demonstrate the most significant impact: the network covariate effect of shared inventor relationships, the network covariate effect of shared patent family relationships, and the transitivity effect. Several auxiliary mechanisms also influence citation relationship formation, including citation lag, number of claims, and number of references.

Patent citations play a crucial role in science and technology evaluation because they can track the evolution of technology, measure technological diffusion and spillovers between countries and regions, assess the quality and value of inventions and technologies, and analyze the technological strategic behaviors of innovation actors [1-2]. In recent years, scholars have introduced network analysis methods into patent citation analysis, yielding numerous patent citation network-related research findings that have greatly enriched the perspective of patent citation analysis. These studies have moved beyond traditional approaches that relied solely on patent citation frequency and instead employed visualization and descriptive statistical methods to examine the structural and

dynamic characteristics of patent citations [3-4].

However, research on the formation mechanisms of patent citation relationships remains a relatively weak link in current studies. The reasons for this gap manifest primarily in two areas: First, limitations in observational perspective. The formation of patent citation networks is a complex process with influencing factors that may include the self-evolution of the patent citation network itself, the attribute characteristics of patents, and external network factors. Using either attribute-based indicators or network indicators alone struggles to adequately explain the formation mechanisms of patent citation relationships [5]. Additionally, research conclusions that hold true from a single perspective may conflict when observed at a higher level. Second, inadequacies in statistical inference methods. Traditional statistical inference methods such as regression analysis are based on attribute data and independence assumptions, which are inappropriate for network analysis where the core object is relational data [6]. While some specialized network data statistical inference methods like the Quadratic Assignment Procedure (QAP) can accommodate the characteristics of network data, they are constrained by their framework and lack extensibility in incorporating different data types [7-8]. These two limitations necessitate exploration of new methods to address the question of patent citation relationship formation mechanisms.

The Exponential Random Graph Model (ERGM) is a research method focused on tie formation [9]. Based on relational data and conditional dependence assumptions, ERGM selects local network structures as network statistics to observe the overall structural characteristics of complex networks, thereby obtaining a comprehensive understanding of network complexity, interdependence, and randomness [8]. This method can overcome the aforementioned limitations in studying patent citation relationship formation mechanisms, enabling researchers to achieve a more comprehensive understanding of these mechanisms. The research objective of this paper is: Under the guidance of tie formation theory, to establish multiple exponential random graph models based on the primary mechanisms that may influence the Nelarabine drug patent citation network. Through testing the network statistical effects corresponding to various mechanisms, this study helps understand which mechanisms affect the formation of the Nelarabine drug patent citation network and the magnitude of their effects.

The article is organized as follows: Section 2 describes the basic process of exponential random graph modeling, briefly outlining the five mechanisms affecting patent citation relationship formation and how these are transformed into corresponding local network configurations and network statistics. Section 3 introduces the experimental data—the Nelarabine drug patent citation network—and the process of using statistical methods to visually demonstrate the statistical characteristics related to the five mechanisms. Section 4 presents model analysis, including model comparison, diagnostics, and goodness-of-fit evaluation. Section 5 provides conclusions and discussion, further examining

the impact of the five core mechanisms on patent citation relationship formation, identifying which mechanisms significantly influence citation relationship formation, which are most important, and the implications for future drug development.

2 Patent Citation Relationship Formation and ERGM

2.1 Mechanisms Influencing Patent Citation Relationship Formation

Scholars have conducted extensive research on the mechanisms of patent citation relationship formation, particularly regarding the structural characteristics of patent citation networks [4, 10-11]. A representative study is Professor A. B. Jaffe's 2017 review of patent citation research progress, which identifies three primary perspectives: measuring invention attribute characteristics such as impact and originality; tracking knowledge flows between individuals, institutions, and regions; and mapping innovation network landscapes. From a tie formation perspective, these studies can be understood as identifying three categories of factors influencing patent citation relationship formation: the attributes of patents themselves, the self-organizational processes of patent citation networks, and the processes by which citation networks are influenced by external factors [12].

This paper adopts a tie formation theory perspective and, through reviewing relevant literature, distills five categories of mechanisms influencing patent citation relationship formation. These mechanisms consist of two components: factors affecting patent citation relationship formation and the effects of these factors on such formation. It should be noted that these five mechanisms are not mutually exclusive and can be expanded based on research needs in the future.

Mechanism 1: Main Effects of Patent Attributes. Main effects primarily measure the influence of node attributes on tie formation. This study focuses on two patent attribute characteristics: the number of patent claims and the number of patent references. Related research suggests that the number of claims reflects the boundary of technological exclusivity, while references reveal the patent's dependence on prior art [2]. Current studies have demonstrated that the number of claims positively influences citation frequency [13], and similarly, the number of references also positively affects citation frequency [4]. Unlike standard statistical analysis, ERGM models focus on relationships between pairs of patents. Therefore, the statistics measured are aggregated values of attributes for patent pairs rather than individual patent attribute values.

Mechanism 2: Difference Effects of Patent Citation Timing. Previous research has demonstrated that patent citations exhibit cohort effects, where the number of citations increases over time. Patent citation lag is commonly used to measure the technological lifecycle, explaining the speed of innovation or technological development. Related studies show that patents tend to cite newer patents, manifested as shorter citation lags being more likely to form

citation relationships [14-15]. Specifically in terms of network effects, patent citation lag can be expressed as the influence of the difference in grant years between citing and cited patents on citation relationship formation.

Mechanism 3: Convergence Effects of Patent Citation Relationships. Forward citation count (citation frequency) has long been a research focus because it reflects a patent's subsequent technological influence [16-17]. In citation networks, the convergence effect of patent citation relationships 刻画 s the distribution of forward citation counts at the network structure level. It treats highly cited patents as local network configurations with star-like structures (two or more arcs linking into the central node) to observe how such configurations influence network tie formation, such as "rich club" [18] or "preferential attachment" [19] phenomena. These are precisely what convergence effects measure. Therefore, Mechanism 3 refers to the influence of convergence structures formed between patent pairs on citation relationship formation.

Mechanism 4: Transitivity Effects of Patent Citation Relationships. Transitivity effects primarily observe a special local network configuration—transitivity closure. Transitivity closure is a pure network structural characteristic that early research mainly measured through indicators such as clustering coefficients. In citation networks, transitivity closure manifests in two aspects: First, it represents an arc added to a 2-path structure, making the "missing link" explicit and rendering the internal relationships more robust. This feature can be used to analyze the evolutionary paths of patented technologies [20]. Second, in-degree distributions in transitivity closure structures are not uniform, with certain nodes having more indegree. This indegree advantage in transitivity closure structures is superior to that in simple convergence effect structures. Thus, transitivity effects can also identify sources in knowledge flow processes [21-22]. Therefore, Mechanism 4 refers to the influence of transitive structures formed between patent pairs on citation relationship formation.

Mechanism 5: Network Covariate Effects of Patent Citation Networks. Unlike the above mechanisms, network covariate effects do not refer to internal network structural characteristics of the patent citation network but rather examine the covariation between other networks and the patent citation network. Existing research reveals correlations between geographic proximity of patent assignees and inventors and patent citation relationship formation [23]. H. D. White's research further confirms that patent citation networks result from the joint action of two network structural features: social structure and intellectual structure [24]. Additionally, semantic similarity between patent texts can also influence citation relationship formation to some extent [25]. Therefore, Mechanism 5 refers to the influence of relationships in other networks on patent citation relationship formation between patent pairs.

2.2 From Influence Mechanisms to Local Network Configurations

ERGM is a research method focused on tie formation, originating from the social network statistical analysis model proposed by P. Erdos and A. Renyi in 1959. In 1996, S. Wasserman expanded this model into the ERGM/p* model that could incorporate any statistical configuration in a graph. In 1999, J. Anderson's parameter estimation method for the model represented significant progress [26]. ERGM is an extensible model that can be adjusted according to research content. Its most general form is:

Pr(Y = y) = (1/) exp{ _A g_A(y)} (Formula 1)

where the summation includes all configurations A, _A is the parameter corresponding to configuration A (which can be used to determine the influence of specific network statistics in the observed network), g_A(y) = y_{ij} A y_{ij} is the network statistic for the corresponding configuration, and is a normalization constant ensuring the formula represents a proper probability distribution [27]. Simply put, the core task of ERGM is to assign weights to networks with certain specific mechanism combinations. Therefore, the above can also be written in a conditional Logit form:

Logit(P(Y_{ij} = 1 | npatent, Y_{ij}^C)) = _A δg_A(y) (Formula 2)

where Y_{ij}^C represents link relationships in the network other than Y_{ij}, and δg_A(y) represents the change in g_A when link Y_{ij} changes from 0 to 1. Thus, Formula (2) means the probability of predicting a new link's appearance under the condition that other links in the network are already determined.

ERGM theoretically solves the evaluation problem of mixed variables (containing multiple attribute and relationship variables) in complex network conditions that traditional methods cannot address. It can explain the causes of patent citation relationships at the whole-network level, thereby enabling more accurate predictions. Table 1 demonstrates the process of transforming the five mechanisms affecting patent citation relationship formation into quantifiable network statistics.

3 Data Sources and Exploration

3.1 Data Sources

This study focuses on a small-molecule innovative anticancer drug whose Chinese name is Nelarabine (奈拉滨) and whose U.S. trade name is Arranon (阿仑恩). Nelarabine was selected based on the following considerations:

The drug development process involves multiple stages, with a complete lifecycle typically spanning 15-20 years. Certain drug-specific characteristics—such as technology transfer in early-stage drug development, clinical trial results from Phases I, II, and III, potential indications, technology transfer during commercialization (financing, mergers and acquisitions), negative reports on toxicity

and efficacy in post-commercialization stages, drug replacement and upgrading, and "patent cliffs" resulting from patent expiration—may potentially affect the scale and characteristics of a drug's patent citation network [28-30], potentially introducing bias into the citation relationship formation patterns discovered by ERGM models. Therefore, in selecting a drug case, we aimed to start with a relatively simple drug. We considered three criteria for screening a simple drug: relatively straightforward development and production processes with minimal involvement in mergers or transfers; narrow range of indications with relatively few negative toxicity and efficacy reports; and observation period extending through a period after the expiration of core patents to enable comprehensive observation of the patent citation network.

The early-stage drug development of Nelarabine was primarily conducted through collaboration between the U.S. National Cancer Institute and GlaxoWellcome. GlaxoWellcome and SmithKlineBeecham merged in 2000 to form GlaxoSmithKline. Although the early drug development stage featured multi-party participation, since core patents were generated after 2005 and citations appeared after 2006, the multi-participant characteristic of early drug development activities does not significantly affect citation relationship formation in this study [31]. Additionally, although the Nelarabine drug development and production team was acquired by Novartis after 2015, because it was a complete acquisition with the drug's development and production remaining entirely under the control of the original GlaxoSmithKline team, this also does not significantly affect final citation relationship formation.

Nelarabine was approved by the FDA in October 2005 as an orphan drug (for treating rare diseases) through the FDA's special approval process [32]. The drug's indication is for the treatment of T-cell acute lymphoblastic leukemia (T-ALL) and T-cell lymphoblastic lymphoma (T-LBL) in patients who have failed at least two treatment regimens or experienced relapse after treatment [33], representing a limited range of potential indications. According to relevant literature, it is an effective drug for treating T-cell malignancies, with all phases of clinical trials achieving good results. The main challenge involves controlling the risk of neurotoxicity through dose adjustment. Later research has primarily focused on combination therapy, and according to current studies, no fully substitutive new drug for this medication has yet emerged [34].

Using the PubChem Compound Database (PubChem Compound Database) for retrieval, with the search strategy selecting Nelarabine's PubChem CID 3011155 (https://pubchem.ncbi.nlm.nih.gov/compound/3011155), relevant patent information can be obtained [35]. As of December 31, 2017, the search results showed 3,035 patent-related documents. In the data preprocessing stage, this study limited the scope to citation relationships between U.S. patent grants from 1998 to 2016. The final dataset included 1,165 U.S. patent grants related to Nelarabine drug compounds and 1,168 patent citation relationships. For data supplementation, the PatentsView patent database (http://www.patentsview.org/api/doc.html) and the USPTO patent full-text

and image database (http://patft.uspto.gov) were used. After data supplementation, the dataset was further processed into network data format, consisting of two datasets: patent attribute data and inter-patent relationship data.

(1) Patent Attribute Data. Patent attribute data contains four fields, where Patent_{id} is the patent identifier, and the other three fields are attribute information related to the patent: patent grant year, number of patent claims, and number of patent references. For data standardization purposes, the number of claims and references were processed (see Table 2).

(2) Inter-Patent Relationship Data. Relationship data (see Table 3) contains five fields, where patent_{{{id}}*{{ego}}} and patent*{{{id}}_{{alter}}} represent the two ends of a relationship. Since patent citation relationships are directed, all other relationship types were converted to directed relationships for processing. The relationship data includes three types of relationships: shared applicant relationships between patents, patent citation relationships, and shared patent family relationships between patents.

3.2 Data Analysis

Before statistical modeling, it is necessary to observe the data using graphical visualization and descriptive statistics. Related research has found that real networks often exhibit many structural differences from random networks, which help distinguish real networks from simple random networks. Through basic data exploration, we discovered that the Nelarabine drug patent citation network demonstrates network structural effects different from random networks across all five mechanisms:

(1) Main Effect Characteristics of Patent Attributes. Table 4 presents the number of patent claims for citing and cited patents in patent citation pairs, establishing a confusion matrix to statistically examine all possible combinations of patent citation pairs with different claim counts and test whether citation relationship formation is influenced by claim count attribute characteristics. In Table 4, columns represent citing parties in patent citation pairs, while rows represent cited parties. The data density is notably higher in the upper-left matrix block (rows 1-5 and columns 1-5), suggesting that patents with fewer claims have a higher probability of establishing citation relationships. Further observation reveals that in the upper-left matrix block, the upper triangle area above the diagonal shows significantly higher data density than the lower triangle area, indicating that patents with lower claim counts are more likely to be cited. Of course, this conclusion requires testing through modeling. Similar main effect characteristics were also observed for patent references.

(2) Temporal Characteristics of Patent Citation Pairs. Since the overall citation span is long but both citing and cited patents were relatively few in the early stages of the Nelarabine drug citation network, Table 5 only displays citation relationships from 2007 to 2016 (1,076 citation relationships) to more clearly demonstrate citation lag characteristics. By establishing a confusion

matrix based on grant years of citing and cited patents, the temporal characteristics between granted patents can be clearly observed. Table 5 reveals two points: First, patent citing for this drug (row sums in Table 5) began to surge after 2013, with only 26 citations in 2012 but 113 citations after 2013, indicating that the drug gradually became a hot topic in drug development research after 2013. Second, there is a high-density region in the 2-5 year range around the diagonal of the adjacency matrix, suggesting that patent citation relationships tend to form between patents with grant time intervals of 2-5 years.

(3) Overall Sparsity and Local Clustering Characteristics of the Patent Citation Network. First, the network density is only 0.000867, indicating that the network is relatively sparse overall. Figure 1 [Figure 1: see original paper]a shows the overall sparse characteristic of the original Nelarabine drug patent citation network, where no highly centralized nodes exist, and the influence range of some high-density regions is limited. Figures 1b and 1c display images after scaling node sizes based on Authority and Hub scores [36] [37]. By comparing the three images in Figure 1, we observe that in local high-density regions of the network, some patents cite and are cited very frequently, demonstrating local clustering characteristics.

(4) Network Covariate Mechanisms. First, as shown in Figure 2 [Figure 2: see original paper], we compare three patent citation networks: the patent citation network, the shared inventor patent citation network, and the shared patent family patent citation network. Figure 2 reveals some basic characteristics of the three networks, particularly Figure 2c, where the shared inventor patent citation network shows a tightly connected core component, indicating the existence of a "small circle" with high self-citation tendency in the Nelarabine drug patent citation network. Within this "small circle," any patent citation pair with a citation relationship shares at least one inventor (i.e., the two patents have a shared inventor relationship). Since high self-citation characteristics reflect the drug technology's dependence on existing technologies to some extent and the important influence of core R&D teams on citation relationship formation, this high self-citation characteristic becomes an important focus for subsequent statistical inference.

While the shared inventor patent citation network presents a relatively clear core component, the shared patent family patent citation network appears very chaotic, with almost no network structural features discernible from Figure 2b. However, when further calculating the autocorrelation matrices of the three networks (see Table 6), the patent citation network and the shared patent family patent citation network show a 0.301 autocorrelation, while the patent citation network and the shared inventor patent citation network show only a 0.193 autocorrelation. Particularly noteworthy is that the distribution of relationship quantities across the three networks is not uniform: the shared inventor patent citation network contains 10,643 relationships, while the shared patent family patent citation network contains only 472 relationships. Combining these two pieces of information, it is evident that some high degree of covariation exists

between the patent citation network and the shared patent family patent citation network. This covariation is unrelated to structural features and may suggest that some strong rules or business logic influence patent citation formation, a judgment that also requires confirmation through network statistical inference.

Another noteworthy issue is how to identify network covariate effects during the data exploration stage. Discovering which networks have covariation with the patent citation relationship network is not an intuitive process. One experience from our early exploration of patent citation attribute features was that some attributes highlighted in existing literature, such as inventor count and patent family size, did not show significant effects. Therefore, instead of directly using whether patent citation pairs share attributes (main effects, differences, homophily, or heterophily) as statistical features, we constructed networks by observing whether citation pairs share inventors or patent family relationships, then measured the covariation between these two networks and the patent citation network. This transformation process often yields unexpected results.

4 Results Analysis

4.1 Parameter Estimation

Model evaluation is an important component throughout the modeling process. Generally, the ERGM research process involves: first using a null model (randomly generated network) as a baseline, then gradually adding network statistics corresponding to different mechanisms to form new models, using ERGM to estimate parameters for these models, and finally conducting diagnostics, goodness-of-fit evaluation, comparison, and interpretation of multiple models. This study uses R's statnet package to estimate parameters for the models in Table 7 . The null model, main effects model, difference model, and covariate model all employ maximum likelihood estimation, while the geometrically weighted model uses Markov Chain Monte Carlo Maximum Likelihood Estimation (MCMC-MLE) [38].

Table 7 presents a statistical summary of five models. By comparing these summaries, particularly analyzing parameter estimates and their statistical significance, preliminary statistical observations of network statistics can be obtained. "Sum of claims for citing patents" shows significant negative effects across all models, indicating that, all else being equal, larger combined claim counts for a patent pair are associated with lower probability of citation relationship formation. "Sum of references for citing patents" shows significant positive effects across all models, indicating that, all else being equal, larger combined reference counts for a patent pair are associated with higher probability of citation relationship formation. Notably, "Sum of references for cited patents" shows significant positive effects in all models except the geometrically weighted model. A possible explanation is that when geometrically weighted indegree distribution or geometrically weighted edge-wise shared partners statistics are added to the model, some degree of correlation may exist among these three factors.

In the difference model, the first two terms "Citation lag (2 years)" and "Citation lag (3 years)" show significant positive effects across all models, indicating that if the grant time interval between patent pairs does not exceed 3 years, they have higher probability of establishing citation relationships. The latter two terms "Citation lag (4 years)" and "Citation lag (5 years)" warrant attention: when geometrically weighted indegree distribution or geometrically weighted edge-wise shared partners statistics are added, "Citation lag (4 years)" becomes non-significant. A possible explanation is that geometrically weighted indegree distribution or edge-wise shared partners statistics correlate with citation lag (4 years). Both "Shared patent family relationship" and "Shared inventor relationship" show significant positive effects in both the covariate model and geometrically weighted model, indicating that, all else being equal, when a shared patent family relationship or shared inventor relationship exists between a patent pair, the probability of establishing a citation relationship increases. Notably, the parameter values for "Shared patent family relationship" and "Shared inventor relationship" are very high (2.8, 2.9) and (6.34, 5.41) respectively, indicating these two network covariate mechanisms have substantial positive impact on citation relationship formation. Finally, "Geometrically weighted indegree distribution" is significantly negative, indicating that the probability of establishing citation relationships between patent pairs is lower than random occurrence, while "Geometrically weighted edge-wise shared partners" is significantly positive, indicating higher probability than random occurrence. While these appear contradictory, together they further illustrate the coexistence of overall sparsity and local clustering in network structure.

4.2 Model Diagnostics

Model diagnostics can assist in determining whether the estimation algorithm has converged or exhibits near-degeneracy problems, thereby judging whether the model itself or model evaluation settings require adjustment [39]. Figure 3 [Figure 3: see original paper] shows the status of some statistics in the geometrically weighted model during the final iteration stage. The left-side plots use MCMC chains to display time series changes for each statistic, while the right-side plots show the distribution of the corresponding MCMC chains. If the model converges, the plots for each statistic will show random variation centered at 0, where 0 represents the statistic value for the observed network. In the geometrically weighted model, most statistics show random variation around 0, indicating that the model diagnostics confirm the geometrically weighted model is stable.

4.3 Model Fit

Although some network statistics have already demonstrated statistical significance in the parameter estimation stage and reflect patterns consistent with preliminary exploratory analysis, providing initial validation of model effectiveness, more systematic testing is needed: to what extent can the simulation

model reflect the structural characteristics of the observed network? We evaluate model goodness-of-fit from two aspects:

(1) Goodness-of-fit evaluation using AIC and BIC statistics. AIC and BIC methods are based on log-likelihood estimation results, measuring the difference between the probability of Y_{ij} (actual observed ties) and the expected probability of Y_{ij} in the observed network. According to Table 7, the null model's AIC is 18,806, the main effects model's AIC is 17,640, showing substantial improvement over the null model. The difference model's AIC is 17,510, showing slight but not significant improvement over the main effects model. The covariate model shows a significant decrease compared to the previous difference model, with AIC dropping to 8,515, indicating that the two mechanisms in the covariate model—shared patent family and shared inventor relationships—play important roles in improving ERGM fit.

(2) However, AIC and BIC methods are suitable for observation data based on independence assumptions. When models become more complex, such as the geometrically weighted model with added dependence statistics, simulation-based goodness-of-fit evaluation methods are needed. The goodness-of-fit evaluation process can also use visual graphical observation methods. By fixing other network characteristics, we compare the log-odds ratio of each parameter in the observed network with the range of log-odds ratios in the simulated network. Figure 4 [Figure 4: see original paper] shows the goodness-of-fit evaluation results for the geometrically weighted model simulation network. The black line represents observed results from the patent citation network; gray lines and boxplots represent measurement results for the simulation network at a 95% confidence interval. When the black line falls between gray lines, the simulation network can well represent the structural characteristics of the real patent citation network. Therefore, Figure 4 indicates that the simulation network can basically fit four structural characteristics of the real network (indegree centrality, outdegree centrality, edge-wise shared partners, and dyad-wise shared partners), though some differences remain in edge-wise shared partners.

4.4 Model Interpretation

By comparing multiple statistical results and graphical goodness-of-fit indicators across five models (null model, main effects model, difference model, covariate model, and geometrically weighted model), we find that the geometrically weighted model has the best network simulation effect. During model construction, we observed three major improvements in network simulation effectiveness: (1) Adding covariate effect statistics, treating networks of shared patent family relationships and shared inventor relationships between patents as network covariates to predict citation relationship formation probability; (2) Adding main effects statistics, considering the influence of attribute factors on citation relationship formation probability, including the number of patent claims and references; (3) Adding geometrically weighted statistics, considering the influence

of indegree distribution and edge-wise shared partners on citation relationship formation probability.

Specifically, interpreting from the ERGM goodness-of-fit improvement effects, the "Shared inventor relationship" statistic has the greatest impact on patent citation relationship formation. The network covariate effect represented by the "Shared inventor relationship" statistic is similar to the "author self-citation" effect in literature, indicating that the Nelarabine drug patent citation network revolves around a "small circle" with high self-citation tendency. This small circle simultaneously occupies hub and authority positions in the network (see Figures 1 and 2), and thus its development determines the overall shape of the Nelarabine drug patent citation network.

The "Shared patent family relationship" statistic reveals another important dimension influencing patent citation relationship formation: the business logic behind patent applications. Patent applicants use derivative patent applications for patent portfolio strategies, such as fencing strategies to expand patent protection periods. This rule is a strong business rule that cannot be discerned from network structural features (see Figure 2b and Table 6) but profoundly influences patent citation relationship formation.

The transitivity effect represented by "Geometrically weighted edge-wise shared partners" is similar to "a friend of a friend is also a friend." For citation networks, the existence of transitivity effects is not difficult to understand. More noteworthy is that in the geometrically weighted model, adding the "Geometrically weighted edge-wise shared partners" statistic causes the relative influence of other statistics to decline, indicating some degree of substitution effect. This is precisely the advantage of ERGM: its ability to provide statistical inference for multiple variables with complex nested relationships, a task that traditional regression models cannot accomplish.

Of course, main effects of patent attributes also play a role. For example, the two patent attribute main effect mechanisms—"Sum of claims for citing patents" and "Sum of claims for cited patents"—indicate that in the Nelarabine drug patent citation network, patents tend to cite patents with fewer claims, adopting a strategy of actively avoiding competitors' claim scopes [30]. Simultaneously, patents tend to cite patents with more references, adopting an active information disclosure strategy to avoid disadvantageous positions in later litigation due to incomplete information disclosure [14].

Difference effects of patent citation timing also play a role. "Citation lag (2 years)" and "Citation lag (3 years)" both show significant difference effects, which is common in patent citations. However, it is important to note that "Citation lag (4+ years)" becomes non-significant after adding transitivity and convergence effect mechanisms, indicating some substitution effect between network structural features (such as transitivity) and difference effects. For example, if three patents form a transitive triad through citation relationships, a patent pair in the triad has both direct and indirect citation relationships. In this case,

citation lag is typically longer than for patent pairs with only direct citation relationships. A reasonable explanation is that although some patent pairs in the patent citation network have long citation lags, these citation pairs often simultaneously exhibit transitive triad structures. Therefore, after considering transitivity, the citation lag (4+ years) statistic is no longer significant.

5 Conclusions and Limitations

This study employs a novel statistical inference method—the Exponential Random Graph Model—which provides a unique perspective enabling comprehensive evaluation of mixed variables under complex network conditions, thereby explaining patent citation relationship formation at a broader level. In terms of micro-structural feature design, this paper considers five mechanisms: main effects, difference effects, covariate effects, convergence effects, and transitivity effects. These mechanisms encompass internal self-organizational structural features of networks, external network covariation, and internal patent attribute features—multiple relationships and highly nested local structures that traditional regression models based on independence assumptions cannot handle.

The main conclusions are as follows: For the Nelarabine drug citation network, patent citation relationship formation is primarily influenced by three factors: The covariation between shared inventor relationships and patent citation relationships reveals a "small circle" with high self-citation tendency that significantly influences the overall direction of Nelarabine drug development. The covariation between shared patent family relationships and patent citation relationships demonstrates the role of business logic behind patent applications—using patent families for portfolio strategies. The internal self-organizational network characteristics of patent citation networks—such as transitivity—show that citation relationship formation is not a random process but exhibits strong dependence on existing network structures.

Additionally, several auxiliary factors influencing patent citation relationship formation in the Nelarabine drug patent citation network merit attention: The substitution effect of network structural features (such as transitivity) on citation lag (4+ years); patents' tendency to cite patents with fewer claims, adopting a strategy of actively avoiding competitors' claim scopes; and patents' tendency to cite patents with more references, adopting an active information disclosure strategy to avoid disadvantageous positions in later litigation due to incomplete information disclosure.

This study has certain limitations worth future exploration. First, this research focuses on a specific drug, so all explanations for patent citation relationship formation discovered currently apply only to the Nelarabine drug patent citation network and lack generalizability—a limitation determined by the characteristics of ERGM as a generative model. However, if we can analyze multiple drugs simultaneously, it may still be possible to induce some universal explanations for patent citation relationship formation. Comparative studies across multiple

drugs represent an important future research direction. Second, this study only observes patent citation relationship formation from a cross-sectional data perspective. Future research observing patent citation relationship formation from a dynamic perspective represents another important research direction.

References

[1] OECD. OECD patent statistics manual[M]. Paris: OECD Publishing, 2009.

[2] JAFFE AB, DERASSENFOSSE G. Patent citation data in social science research: overview and best practices[J]. Journal of the Association for Information Science and Technology, 2017, 68(6): 1360-1374.

[3] YANG GC, LI G, LI CY. Using the comprehensive patent citation network (CPC) to evaluate patent value[J]. Scientometrics, 2015, 105(3): 1319-1346.

[4] VAN RAAN AFJ. Patent citation analysis and its value in research evaluation: a review and a new approach to map technology-relevant research[J]. Journal of data and information science, 2017, 2(1): 545-538.

[5] MORRISSA, VANDERVEERMARTENS B. Mapping research models for management research: a case study of executive recruitment[J]. European management journal, 2017, 35(3): 373-385.

[6] ARRIETAPAREDES MP, CRONIN B. Exponential random graph models for social networks[J]. Social networks, 2007, 29(2): 173-191.

[7] ROSEKIM JY, HOWARD M, COXPAHNKE E. Understanding network formation in strategy research: exponential random graph models[J]. Strategic management journal, 2016, 37(1): 22-44.

[8] GOODREAU SM, HANDCOCK MS, BUTTS CT. Statnet: software tools for the representation, visualization, analysis and simulation of network data[J]. Journal of statistical software, 2008, 24(1): 1-11.

[9] ROBINS G, PATTISON P, KALISH Y. An introduction to exponential random graph (p*) models for social networks[J]. Social networks, 2007, 29(2): 173-191.

[10] JAFFE AB, TRAJTENBERG M. Patents, citations, and innovations[M]. New York: MIT Press, 2002.

[11] ALCÁCER J, GITTELMAN M. Patent citations as a measure of knowledge flows: the influence of examiner citations[J]. Review of economics and statistics, 2006, 88(4): 774-779.

[12] ROBINS G. Doing social network research[M]. London: SAGE, 2015.

[13] FISCHER T, LEIDINGER J. Testing patent value indicators on directly observed patent value: an empirical analysis of Ocean Tomo patent auctions[J]. Research policy, 2014, 43(3): 519-529.

[14] ALCÁCER J, GITTELMAN M, SAMPAT B. Applicant and examiner citations in U.S. patents: an overview and analysis[J]. Research policy, 2009, 38(2): 415-427.

[15] HALL BH, JAFFE AB, TRAJTENBERG M. The NBER patent citation data file: lessons, insights and methodological tools[R]. Cambridge: National Bureau of Economic Research, 2001.

[16] BENSON CL, MAGEE CL. Quantitative determination of technological improvement from patent data[J]. Public library of science, 2015, 10(4): e0121635.

[17] CZARNITZKI D, HUSSINGER K, SCHNEIDER C. "Wacky" patents meet economic indicators[J]. Economics letters, 2011, 113(2): 131-134.

[18] SMILKOV D, KOCAREV L. Rich-club and page-club coefficients for directed graphs[J]. Physica a: statistical mechanics and its applications, 2010, 389(11): 2290-2299.

[19] BRANTLETF, FALLAH MH. Complex innovation networks, patent citations and power laws[C]//PICMET'07-2007 Portland international conference on management of engineering & technology. Portland: IEEE, 2007: 540-549.

[20] WANG JC, CHIANG CH, LIN SW. Network structure of innovation: can brokerage or closure predict patent quality?[J]. Scientometrics, 2010, 84(3): 735-748.

[21] BATAGELJ V. Efficient algorithms for citation network analysis[EB/OL]. [2017-12-31]. https://arxiv.org/abs/cs/0309023.pdf.

[22] HUNG SW, WANG AP. Examining the small world phenomenon in the patent citation network: a case study of the radio frequency identification (RFID) network[J]. Scientometrics, 2009, 82(1): 121-134.

[23] ALMEIDA P, KOGUT B. The exploration of technological diversity and geographic localization in innovation: start-up firms in the semiconductor industry[J]. Small business economics, 1997, 9(1): 21-31.

[24] WHITE HD, WELLMAN B, NAZER N. Does citation reflect social structure?: longitudinal evidence from the "Globenet" interdisciplinary research group[J]. Journal of the Association for Information Science and Technology, 2004, 55(2): 111-126.

[25] YAN E, DING Y. Scholarly network similarities: how bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other[J]. Journal of the Association for Information Science and Technology, 2012, 63(7): 1313-1326.

[26] SNIDERS TAB, PATTISON PE, ROBINS GL. New specifications for exponential random graph models[J]. Sociological methodology, 2006, 36(1): 99-153.

[27] ROBINS G, SNIDERS T, WANG P. Recent developments in exponential random graph (p*) models for social networks[J]. Social networks, 2007, 29(2):

192-215.

[28] THORNE N, AULD DS, INGLESE J. Apparent activity in high-throughput screening: origins of compound-dependent assay interference[J]. Current opinion in chemical biology, 2010, 14(3): 315-324.

[29] LIM SY, SUH M. Intellectual property business models using patent acquisition: a case study of royalty pharma inc[J]. Journal of commercial biotechnology, 2016, 22(2): 6-18.

[30] WAGNER S, WAKEMAN S. What do patent-based measures tell us about product commercialization? evidence from the pharmaceutical industry[J]. Research policy, 2016, 45(5): 1091-1102.

[31] KISOR DF. Collaboration to meet therapeutic need: the development of nelarabine[J/OL]. Clinical medicine. 2009, 1: 1317-1320. [2018-12-06]. https://dio.org/10.4137/CMT.s2909.

[32] FDA approval for nelarabine[EB/OL]. [2017-12-06]. https://www.cancer.gov/about-cancer/treatment/drugs/fda-nelarabine.

[33] COHEN MH, JOHNSON JR, JUSTICE R. FDA drug approval summary: nelarabine (Arranon) for the treatment of T-cell lymphoblastic leukemia/lymphoma[J]. The oncologist, 2008, 13(6): 709-714.

[34] KADIAT TM, GANDHI V. Nelarabine in the treatment of pediatric and adult patients with T-cell acute lymphoblastic leukemia and lymphoma[J]. Expert review of hematology, 2016, 10(1): 1-11.

[35] PAPADATOS G, DAVIES M, DEDMAN N. SureChEMBL: a large-scale, chemically annotated patent document database[J]. Nucleic acids research, 2016, 44(D1): D1220-D1228.

[36] MARRA M, EMROUZNEJAD A, HO W. The value of indirect ties in citation networks: SNA analysis with OWA operator weights[J]. Information sciences, 2015, 314: 135-151.

[37] LUKE D. A user's guide to network analysis in R[M]. Cham: Springer International Publishing, 2015.

[38] DUBNJAKOVIC A. An evaluation of exponential random graph modeling and its use in library and information science studies[J]. Library & information science research, 2016, 38(3): 259-264.

[39] ROBINS G, PATTISON P, WANG P. Closure, connectivity and degree distributions: exponential random graph (p*) models for directed social networks[J]. Social networks, 2009, 31(2): 105-117.

Author Contributions:

Yang Guancan: Responsible for conceptual framework development, main content writing, and experimental results analysis.
Liu Zhanlin: Conducted data exploration, code optimization, and experimental

results analysis.
Li Gang: Determined research direction and provided paper revision suggestions.

Note: Figure translations are in progress. See original paper for figures.