The Impact of Document Similarity on Scientific Citation Preferences: An Empirical Study (Post-print)

Authors: Duan Qingfeng, Pan Xiaohuan

Date: 2023-08-26T00:00:00+00:00

Abstract

[Purpose/Significance] Citing and cited documents often exhibit certain similarities. Revealing the formation mechanisms behind this phenomenon contributes to a deeper understanding of the nature of citations.

[Method/Process] This study employs exponential random graph models to conduct empirical analysis in the field of library and information science, aiming to reveal the influence mechanism of document similarity on citation relationships.

[Results/Conclusions] Empirical findings indicate a significant tendency toward similarity-based citation at the network structure, institutional, and journal levels. Specifically, citation relationships are more likely to be embedded in triadic transitive structures; documents from the same institutions and journals are more likely to cite each other; and documents from countries with disciplinary dominance are more likely to cite each other. The empirical results fully demonstrate that social proximity is an important formation mechanism of citation behavior, reflecting the social attributes of citation preferences.

Full Text

Preamble

ChinaXiv Cooperative Journal, Vol. 62, No. 4, February 2018

An Empirical Study on the Impact of Document Similarity on Scientific Citation Preferences

Duan Qingfeng[1], Pan Xiaohuan[2]

[1] School of Management Science and Engineering, Shanxi University of Finance and Economics, Taiyuan 030006

[2] School of Economics and Management, North University of China, Taiyuan 030051

Abstract

[Objective/Significance] Citing and cited documents often exhibit certain similarities. Revealing the underlying formation mechanisms of this phenomenon contributes to a deeper understanding of the nature of citations. [Method/Process] This study employs Exponential Random Graph Models (ERGM) to conduct empirical analysis in the field of library and information science, aiming to uncover the influence mechanisms of document similarity on citation relationships. [Results/Conclusions] The empirical findings reveal significant tendencies toward similarity among citing documents at the network structure, institutional, and journal levels. Specifically, citation relationships are more likely to be embedded in transitive triadic structures; documents from the same institutions and journals are more prone to cite each other; and documents from countries with dominant positions in the discipline are more likely to form citation relationships. These results demonstrate that social proximity constitutes an important mechanism in citation behavior formation, reflecting the social attributes of citation preferences.

Keywords: documents; similarity; scientific citation; Exponential Random Graph Model
Classification Number: G253
DOI: 10.13266/j.issn.0252-3116.2018.04.013

In modern scientific development, standardized document citation plays a crucial role and has stimulated scholars' interest in studying its connotations and mechanisms. Notably, citations have become a fundamental theoretical cornerstone and measurement tool for modern academic evaluation, with metrics such as journal impact factors and h-indexes all relying on citation data. However, a critical issue cannot be ignored: evaluation theories and tools based on citations essentially presuppose an idealized premise that the selection of cited objects is entirely based on academic value judgments or certain academic purposes. Clearly, there exists a substantial gap between this strict assumption and practical applications, with various controversies and academic debates reflecting the complexity of citation connotations. Therefore, it is necessary to trace back to the origin of the problem and explore the mechanisms of citation formation. An accurate understanding of citation value may help propose or revise citation-based evaluation indicators with greater academic value discrimination effectiveness.

Given that citations essentially represent dyadic relationships between documents, the similarity between citing and cited documents may serve as an effective perspective and approach for examining the intrinsic mechanisms of citation behavior. This similarity may manifest in various incidental phenomena, and the so-called "academic circles" represent a form of social proximity. Some

scholars have proposed the “elite club” index to identify these highly mutually recognized scholar groups from a statistical perspective, using citation relationships as clues.

Self-citation can be viewed as a unique citation phenomenon where citing and cited documents originate from exactly the same author. Research on self-citation is abundant, particularly focusing on its fairness and applicability issues in scientific evaluation. On one hand, some scholars argue that uncontrolled self-citation quantities distort the normal distribution of citations, questioning the fairness of evaluation indicators that include self-citations. On the other hand, others contend that self-citation cases require specific analysis, and when sample sizes are sufficiently large, there is no need to exclude self-citations. Some questionnaire surveys have even found no significant difference in motivation between self-citation and other-citation among researchers.

The tendency toward document similarity accompanying citation relationships may manifest at multiple levels, including authors, countries or regions, languages, journals, and institutions. For example, A. Bookstein and M. Yitzhaki designed a mother-tongue preference indicator to study citation tendencies among same-language groups, demonstrating the influence of language preference on citation behavior. S. Ren and R. Rousseau conducted empirical research from a journal citation perspective, discovering high mutual citation rates among Chinese journals. Tang Li et al. analyzed the “club effect” behind China’s rapid growth in scientific output, showing through Sino-US comparative analysis that Chinese authors exhibit more significant mutual citation behavior among highly cited papers.

Although the phenomenon of document convergence in citation relationships has attracted some scholarly attention, existing research remains fragmented and lacks systematic investigation, with insufficient theoretical analysis of the underlying mechanisms. Reviewing relevant literature reveals several deficiencies: (1) Theoretical discussions dominate while empirical analyses are scarce, particularly lacking modeling analyses that fully utilize large-scale bibliographic data; (2) Document convergence during citation processes exhibits multidimensional manifestations, yet existing literature mostly focuses on single dimensions without systematic discussion under a unified analytical framework; (3) Most studies employ descriptive analysis without statistical inference, making it difficult to determine the extent to which document similarity actually influences citation preference formation.

Thus, this paper addresses the fundamental question: Is document similarity accompanying citation relationships ubiquitous, and to what extent does it influence citation relationship formation? Analyzing external social factors’ impact on academic citation processes through empirical results will contribute to a more comprehensive understanding of the nature of scientific citations.

Therefore, guided by social construction theory, this paper proposes research hypotheses regarding document similarity tendencies in citation behavior at dif-

ferent levels. Taking the field of library and information science as the empirical domain and employing Exponential Random Graph Models, this study models the probability of citation relationship formation as a function of relevant document similarity indicator variables, ultimately exploring the fundamental mechanisms of citation formation through empirical analysis.

2 Research Methods

2.1 Exponential Random Graph Model

The Exponential Random Graph Model (ERGM) is a statistical analysis model constructed for network dyadic relationships, also known as the p* model. It has become an important and widely applied method for modeling network edges, capable of depicting the influence of various factors on network dyadic relationship formation. The model formulates the probability of observing an actual network y as a function of various possible configurations (e.g., number of edges, triangles, reciprocity) and node/edge attributes. The ERGM is defined as [11]:

$$p(Y = y|X) = \frac{\exp\{\theta^T g(y, X)\}}{x(\theta, y)}$$

where Y is a set of random relationships represented by a random adjacency matrix Y_{ij}, with elements in row i and column j corresponding to the relationship from node i to node j; y is a realization of the random adjacency matrix Y, representing the specific observed relationships; X is a vector of covariates related to edges or nodes; is a vector of coefficients corresponding to various variables; and g(y,X) is a vector of network statistics. If a configuration is observed k times in network y, then g(y) = k. x(,y) is a normalizing factor ensuring that the sum of probabilities for all possible network samples equals 1, i.e., Σ exp{ ^T g(z,X)} = 1.

To further illustrate the interpretability of parameter , we introduce the change statistic δ, as shown in formula (2):

$$\delta g(y)_{ij} = g(y_{ij}^{+}) - g(y_{ij}^{-})$$

where y_{ij} represents the dyadic relationship between nodes i and j, with 1 indicating a connection and 0 otherwise. y^+*{ij} and* y^-{ij} represent network realizations with y_{ij} set to 1 or 0 respectively, while keeping the rest of y unchanged. Thus, δg(y)*{ij} reflects the change in network statistic g(y) when* y{ij} changes from 0 to 1 while other edges remain constant. Using the change statistic, formula (1) can be equivalently transformed into another form, as shown in formula (3):

$$\text{logit}[P_{\theta,y}(Y_{ij} = 1|Y_{ij}^{c})] = \theta^T \delta g(y)_{ij}$$

where the logit function is defined as logit(p) = log[p/(1-p)], P is the probability of a dyadic relationship occurring, and Y^c_{ij} represents the rest of network Y except for edge Y_{ij}. The left side represents the log-odds of nodes i and j forming a connection given that the rest of the network Y^c_{ij} remains unchanged. The right side indicates that, with the rest of the network fixed, when y_{ij} changes from 0 to 1, each unit increase in network statistic g(y_{ij}) makes the odds of i and j forming a connection exp( ) times the odds of not forming one. The magnitude of parameter reflects the marginal effect of various factors on network edge relationships.

This study employs ERGM for modeling analysis based on three main considerations: (1) Documents form citation networks through citation relationships, making ERGM, which is specifically designed for modeling network dyadic relationships, highly suitable for network samples; (2) ERGM is more appropriate for network relationship modeling than traditional regression models as it fully accounts for the autocorrelation among network edges, meeting the requirements of network modeling; (3) ERGM can interpret the probability of citation dyad formation as a function of endogenous configurations and exogenous covariates in the citation network, enabling the inclusion of various similarity variables in the model. Its statistical inference capabilities help identify the degree of influence of different forms of document similarity on citation relationships.

2.2 Research Hypotheses

From the perspective of citation network formation, factors influencing each citation relationship can be broadly divided into two categories: (1) Endogenous structural factors arising from interdependencies among citation relationships due to network edge autocorrelation; and (2) Exogenous factors originating from node or edge attributes.

2.2.1 Endogenous Structural Convergence Tendency

Citation relationships are embedded within the literature network. Examining citation behavior from a network perspective helps reveal its complexity and multifaceted characteristics. Edges in network systems are not independent but rather interdependent and influential, potentially forming stable topological structures. From a network formation mechanism perspective, the edge establishment process may be influenced by such endogenous factors, creating connection tendencies that form certain relationships with higher probabilities to obtain network advantages. Nodes embedded in structural patterns also benefit from network effects. Structural characteristics are a focus of social network analysis, manifesting at different scales of network systems—for example, clustering at the macro level, communities at the meso level, and the simplest yet most basic structure at the micro level: the triadic structure. Simply put, using the classic friendship network as an example, the notion that “a friend of a friend may also be a friend” reflects this basic triadic structure, which has been confirmed in actual networks such as journal citation networks [12] and

collaboration networks [13].

Citation networks are directed networks, and transitive triadic structures represent one fundamental configuration. Generally, if document i cites document j, and document j cites document k, then document i will also cite document k. Embedded in a transitive triadic structure, documents i and k have redundant paths between them—besides a direct connection, there exists an indirect path of length 2. From a knowledge flow perspective, document j acts as a knowledge broker between documents i and k, absorbing knowledge from one party and transmitting augmented knowledge to the other. This third-party document sharing mechanism provides a potential knowledge transfer channel between the two documents, avoiding the costs and risks associated with establishing new knowledge flow channels. Therefore, documents that are not directly connected but share common knowledge have a tendency to establish direct connections, forming transitive triadic structures. From the perspective of knowledge flow efficiency, redundant path structures provide a more efficient knowledge dissemination system, reducing the risk of knowledge transmission chain rupture due to system-level vulnerabilities and enhancing the robustness of knowledge transmission networks. The embeddedness of transitive triadic structures reflects the structural convergence tendency of scientific literature at the micro level of citation networks. Therefore, based on social network theory, we propose:

Hypothesis 1: Scientific citation relationships tend to be embedded in transitive triadic structures.

2.2.2 Exogenous Convergence Tendency

Among the guiding theories of scientific citation formation mechanisms, social construction theory has gained widespread attention. It posits that citation behavior emerges more as a social process influenced by multiple external social factors such as politics and economics. The driving factors for citation relationships are not limited to purely academic domains but possess more complex and diverse social attributes. For example, L. Bornmann and H. Daniel systematically reviewed research on citation motivations, emphasizing the influence of non-academic factors in scientific citation generation [14]. This theory provides explanations for the non-academic associations observed between citing and cited documents, suggesting that examining broader social relationships helps understand the complexity of citations.

If certain social relationships between citing and cited documents are not coincidental but co-occur with citation relationships, there may exist non-negligible intrinsic connections between them. Accurately revealing these dependencies would contribute to a more comprehensive understanding of the nature of citations. From a knowledge flow perspective, new knowledge or ideas more easily flow to document nodes with stronger absorption or reception capabilities, and ideas embedded in similar knowledge social networks are more readily accepted. The more similar two documents are in various social attributes, the more likely

they are to establish explicit citation relationships.

If scientific literature is viewed as the final product of knowledge production, then all elements involved in the knowledge production system are indispensable. Innovation agents are producers, disseminators, and absorbers of knowledge. Knowledge dissemination relies on tangible carriers, and knowledge flow and spillover are constrained by spatial and organizational boundaries. According to social construction theory, citations are not only manifestations of intangible knowledge interaction and collision but also results of intertwined environmental elements. Beyond psychological, intellectual, and knowledge factors, citation preference formation and knowledge flow direction should also be consequences of social factors. Therefore, document similarity tendencies may manifest in the following aspects:

(1) **Carrier Homophily.** As academic knowledge carriers, each journal often exhibits distinct disciplinary domain and topic selection preferences, meaning documents published in the same journal have more similar knowledge structures, while those in different journals do not. Citations are fundamentally external representations of knowledge exchange and derivation [15], and homogeneous knowledge is more easily absorbed and understood. Additionally, there may be cases of improper excessive self-citation between publishers and authors pursuing journal impact factors, though such cases may be isolated, they cannot be ignored. Based on this, we propose:

Hypothesis 2: Documents from the same journal are more likely to form citation relationships.

(2) **Formal Organizational Homophily.** Scientific research has long become a profession, with researchers affiliated with academic institutions. Innovation agents within these institutions naturally form long-term and stable academic relationships. The same academic affiliation implies embedded academic relationships, reflecting institutional, systemic, and organizational arrangements and guarantees. From a social network perspective, strong ties not only serve as paths for explicit knowledge transfer but also provide channels for tacit knowledge spillover through long-term collaboration and face-to-face communication. Based on this, we propose:

Hypothesis 3: Documents from the same author affiliation are more likely to form citation relationships.

(3) **Informal Organizational Homophily.** Today's scientific research collaboration and exchange trends are increasingly evident, facilitated by the Internet and social media. "Invisible colleges" have emerged as complementary models to formal academic organizations. Scholar groups, communities, and circles form loose coupling innovation systems based on common or similar fields, interests, tasks, or goals [16]. These informal organizations and groups have become important channels for scholars to acquire new knowledge and expand social relationships. External social

network connections and embeddedness may bring scholars additional heterogeneous knowledge, academic resources, reputation, and advantages. Based on this, we propose:

Hypothesis 4: Documents by authors from the same informal organization are more likely to form citation relationships.

(4) **Geospatial Homophily.** New economic geography has conducted in-depth research on the agglomeration of economic resources, and technological innovation as an endogenous development driver exhibits similar characteristics. The academic community widely recognizes that the geographical clustering of innovation agents facilitates knowledge spillover, and this external effect promotes the efficiency and effectiveness of innovation activities [17]. Beyond knowledge spillover, other social factors may be implicated, such as proximity in geographical space often meaning innovation agents' embeddedness in local social environments, including national politics, legal systems, and language culture. These homogenized external social structures provide a foundation for cognition and identification among innovation agents, enhancing academic exchange effectiveness. Based on this, we propose:

Hypothesis 5: Documents from similar geographical spaces are more likely to form citation relationships.

2.3 Model Specification and Variable Selection

Table 1 presents the configuration variables required in our model, their meanings, and statistical definitions, which are explained below in conjunction with specific variables.

Table 1. Meanings of ERGM Network Statistics

Statistic	Definition	Interpretation
Edges	Σ_{i,j} y_{ij}	Model constant term, equivalent to network density
Ttriple	Σ_{i,k,j} y_{ik}y_{kj}y_{ij}	Transitive triadic structure: Do citation relationships tend to form closed patterns?
Nodematch(δ)	Σ_{i,j} y_{ij}δ_iδ_j	Node homophily: Are documents with identical δ attributes more likely to cite each other?
Nodeicov(δ)	Σ_{i,j} y_{ij}δ_j	Node covariate: Are documents with stronger δ attributes more likely to be cited?

(1) Endogenous Structural Convergence Variables. For directed citation networks, transitive triadic structures as network configurations reflect interdependencies among citation relationships. The variable Ttriple is defined as the number of transitive triads in the citation network, as shown in Table 1. If the coefficient before Ttriple is significantly positive, it indicates that citation relationships are more likely to be embedded in transitive triadic structures compared to other random structures, with larger values indicating stronger tendencies.

(2) Exogenous Convergence Variables. This paper examines four aspects of exogenous convergence tendencies: carrier homophily, formal organizational homophily, informal organizational homophily, and geospatial homophily. As shown in Table 1, if nodes i and j share the same categorical attribute δ, it is recorded as 1, otherwise 0. The Nodematch(δ) statistic represents the number of dyads (i,j) with identical attribute δ in the network, reflecting the degree of attribute matching among nodes.

Carrier homophily is defined by whether citing and cited documents are published in the same journal. The variable Nodematch(JO) is defined as the number of all citing-cited document pairs from the same journal. A significantly positive coefficient indicates that documents from the same journal are more likely to form citation relationships, with larger coefficients indicating stronger marginal effects.

Formal organizational homophily is captured through matching relationships of author affiliations in scientific documents. The variable Nodematch(JG) is defined as the number of all citing-cited document pairs with identical author institutions. Considering that a single document may have multiple authors and each author may belong to multiple institutions, two documents are considered to have the same affiliation if they share at least one author institution. A significantly positive coefficient indicates that documents from the same author institution are more likely to form citation relationships, with larger values indicating stronger marginal effects.

Informal organizational homophily can be captured by whether scientific document authors are similarly highly cited authors. Informal organizations take various forms; here we define them as highly cited authors in the disciplinary field—typically those with high academic influence and prestige who lead the discipline's frontier. Specifically, this study defines authors in the top 1% of cumulative citations in the disciplinary field as highly cited authors. The variable Nodematch(AU) is defined as the number of citing-cited document pairs authored by highly cited authors. A significantly positive coefficient indicates that citation relationships are more likely to occur among documents by highly cited authors, with larger coefficients indicating stronger marginal effects.

Geospatial homophily is captured through matching relationships of document source cities. This study uses whether documents belong to the same city as a proxy variable for geospatial homophily. While less precise than geodesic

distance between cities, this approach is simple, feasible, and sufficient for research purposes. Considering that scientific documents may have multiple city addresses, two documents are considered to be from the same city if they share at least one source city. The variable Nodematch(CY) is defined as the number of citing-cited document pairs from the same source city. A significantly positive coefficient indicates that citation relationships are more likely to occur among documents from the same source city, with larger coefficients indicating stronger marginal effects.

(3) Control Variables. The model includes Nodeicov(CT) to represent academic value, explaining the effect of academic value in the citation process. Essentially, citations represent the interaction and collision of ideas and viewpoints, where the ideological content and academic innovation of documents are crucial to citation impact. The variable Nodeicov(CT) is defined as the cumulative citation count of cited documents, where CT represents the cumulative citation count of a document. This variable tests how the strength of cited document attribute CT affects the probability of connection establishment.

The variable Edges represents the total number of edges in the actual citation network, reflecting the overall volume of citations. This term is mandatory in the model, equivalent to the constant term in traditional regression models, with explanatory power equivalent to network density. The estimated coefficient before this variable reflects the marginal effect of network density on edge connection odds.

The model also introduces the geometrically weighted dyadwise shared partners (GWDSP) variable, defined as a weighted linear combination of the distribution of all possible shared-partner dyads. This variable characterizes the tendency to form open triadic structures in the network while also helping reduce the risk of model degeneracy [18].

3 Empirical Analysis

3.1 Data Sources and Processing

Descriptive analysis at the country, institution, and journal levels helps preliminarily assess the distribution patterns of document similarity phenomena in citations. This study employs the internal citation rate to measure citation convergence degree, defined as the proportion of mutual citations among documents from the same entity (e.g., country, institution, journal) relative to their total citation volume, reflecting the similarity degree of documents in citation relationships. A higher internal citation rate indicates a stronger tendency to select documents from the same entity as citation targets. Additionally, to identify an entity's influence in the discipline, citation counts are used as an indicator—more accumulated citations indicate greater academic influence.

The study selects library and information science (LIS) as the research field, focusing on seven representative international journals: *Journal of Documen-*

tation, *Scientometrics*, *Journal of Information Science*, *Electronic Library*, *Information Technology and Libraries*, *Library & Information Science Research*, and *Journal of the American Society for Information Science*. All data are sourced from the Web of Science database, covering 1980-2010, with document type filtered as "Article," yielding 6,111 initial records. To facilitate analysis, isolated documents in the citation network were removed. Specifically, for any document d in set D, if c C where C is the set of all forward and backward citations of document d, and c D, then document d is defined as an isolated document in the citation network. Following this method, the data were filtered to obtain 2,125 document records.

Model analysis is based on network relationships, requiring extraction of citation relationships from document metadata to form matrix data. Scientific documents are treated as network nodes; if document i cites document j, a directed edge from i to j is formed. The network matrix consists of binary values, with 1 representing a citation relationship and 0 otherwise, resulting in a 2,125 × 2,125 citation relationship matrix. Additional metadata for each paper were also extracted, including source journal, publication year, citation count, authors, institutions, and countries, for generating exogenous covariates in the model.

3.2 Descriptive Analysis

Country and Institution Level Analysis

Table 2 presents internal citation rates by country and institution, listing the top 10 countries and institutions by total citation count.

At the country level, the United States undoubtedly stands out with an extremely high internal citation rate of 70%, reflecting its absolute leading position in the field. Its research essentially represents the disciplinary frontier, forming an internally circulating disciplinary ecosystem from a knowledge flow perspective. The other top 5 countries by influence also show high internal citation rates, distributed in the range of 32%-39%. Other countries in the top 15 (except Brazil) maintain internal citation rates between 10%-20%, with Chinese documents showing a 19% internal citation rate, consistent with overall distribution patterns. Notably, Brazil's 41% internal citation rate significantly exceeds other countries at similar influence levels, suggesting an abnormal knowledge flow structure that may reflect relatively closed research characteristics.

At the institution level, most top 15 institutions by total citation count have high self-citation rates, distributed between 20%-50%. Universities or research institutes such as Leiden University and Drexel University possess strong capabilities in library and information science, with both high internal citation rates and citation volumes. Notably, the Institute for Scientific Information (ISI) has an internal citation rate as high as 55%. Examining its publications reveals that most internal citations relate to pioneering figures in scientometrics such as H. Small and E. Garfield, whose groundbreaking research has become the ide-

ological source and theoretical foundation for current studies. Similarly, many institutions with high internal citation rates can list several highly influential representative scholars—for example, Drexel University has renowned scholars K.W. McCain, H.D. White, and B.C. Griffith, while Leiden University has H.F. Moed, R.J.W. Tijssen, and A.F.J. Van Raan. Stable scholar groups within research institutions facilitate efficient knowledge sharing and flow, contributing to the formation of high-quality research teams and resulting in higher internal citation tendencies.

The data distribution in Table 2 reveals a correlation between entity academic influence (measured by citation count) and internal citation rate at both country and institution levels. To further examine this correlation, Table 3 lists internal citation rate distributions by citation count ranking. For example, countries and institutions in the top 1% by citation volume have average internal citation rates of 70% and 34%, respectively; those in the top 50% have rates of 34% and 18%; and the overall averages are 32% and 15%. As academic influence decreases, internal citation rates show a gradient decline.

At the country level, the Spearman rank correlation coefficient between internal citation rate and citation count is 0.785, significant at the 1% level (two-tailed test). At the institution level, the correlation is 0.493, also significant at the 1% level. This indicates a positive correlation between internal citation rate and citation count, with high-influence countries and institutions typically exhibiting higher internal citation rates.

Journal Level Analysis

Table 4 presents internal citation rate distributions at the journal level. Overall, different journals show substantial variation in internal citation rates. *Scientometrics* and *Journal of the American Society for Information Science* have the highest internal citation rates and demonstrate clear advantages in both citation count and impact factor, reflecting their high influence in the discipline. According to the "preferential attachment" mechanism in complex network theory, these journals have higher impact factors and greater advantages and probabilities of being cited by other documents compared to other journals. Other journals have lower internal citation rates, reflecting slightly inferior academic influence in terms of both impact factor and citation count.

3.3 ERGM Model Analysis Results

The magnitude and significance level of model parameters serve as the basis for analyzing the influence degree of various configuration variables on the dyadic dependent variable. Parameter estimation was conducted using the STATNET package in R environment, specifically employing Markov Chain Monte Carlo Maximum Likelihood Estimation (MCMC-MLE). To assess parameter fitting effectiveness, t-statistics were used for significance testing. Additionally, AIC and BIC indicators serve as criteria for overall model fit.

A stepwise variable addition strategy was adopted for model specification and selection. ERGM parameter estimation results are presented in Table 5 . Model 1 includes only endogenous factors, while Model 2 adds exogenous factors. In Model 2, the homophily variable Nodematch(AU) did not pass significance testing and was therefore removed, forming Model 3. Compared with other models, Model 3 has the smallest AIC and BIC values, with all statistical parameters significant at the 0.1% level, indicating appropriate model specification and satisfactory parameter fitting. The following analysis focuses on Model 3.

Table 5. ERGM Parameter Estimation Results

Variable and Indicator	Model 1	Model 2	Model 3
Endogenous Structural Convergence			
Edges	-7.298 (0.053)***	-7.692 (0.034)***	-7.763 (0.074)***
Transitive Triad Ttriple	2.076 (0.001)***	1.892 (0.001)***	1.947 (0.000)***
GWDSP	-0.131 (0.018)***	-0.115 (0.010)***	-0.167 (0.015)***
Exogenous Convergence			
Journal Nodematch(JO)	-	1.314 (0.055)***	1.285 (0.062)***
Institution Nodematch(JG)	-	3.440 (0.052)***	3.537 (0.157)***
Highly Cited Author Nodematch(AU)	-	-0.005 (0.031)	-
Country Nodematch(CY)	-	-1.123 (0.068)***	-0.852 (0.094)***
Control Variable			
Citation Count Nodeicov(CT)	-	0.012 (0.001)***	0.006 (0.001)***
AIC	52,847.65	52,799.52	52,798.63
BIC	52,867.89	52,835.71	52,819.84

Note: Values in parentheses are standard errors of parameter estimates; , , represent $p < 0.001$, $p < 0.01$, $p < 0.05$, respectively.

The significantly positive coefficient for variable Ttriple indicates that, holding other parts of the network constant, the odds of a citation relationship being embedded in a transitive triadic structure are 7 (= e^1.947) times those of other configurations. If two documents have an indirect citation path, they are more likely to establish a direct citation relationship, demonstrating structural convergence tendency at the citation network level. Furthermore, the positive effect also reflects a tendency for citation relationships to form closed triads, where edges are conditionally dependent on each other and each citation relationship is embedded in a network environment. This empirical result supports Hypothesis 1.

The significantly positive coefficient for variable Nodematch(JO) indicates that journal homophily promotes citation relationship formation. Holding other network conditions constant, the odds of citation relationships occurring between documents from the same journal are 3.61 (= e^1.285) times those between documents from different journals. On one hand, similar research domains and topics facilitate the exchange of ideas and knowledge, while the high visibility

of the same journal carrier also increases citation opportunities. Qiu Junping et al. [19] used questionnaire surveys to explore five motivational psychological structures and factors for researchers' paper citations, finding the important role of information source convenience in citation motivation, which is greatly enhanced by the high visibility of same-journal sources. On the other hand, according to social construction theory, there may be ethical risks of interest exchange between submitting authors and publishers during the publication process, and potential citation suggestions or tacit understandings may also drive up intra-journal citation proportions. This empirical result supports Hypothesis 2.

The significantly positive coefficient for variable Nodematch(JG) indicates that institutional homophily promotes citation relationship formation. Holding other network conditions constant, if two documents share the same institution, the odds of them establishing a citation relationship are 34.36 (= e^3.537) times those of documents from different institutions. Notably, the effect of institutional homophily is exceptionally strong, exceeding the heterophily case by 34 times. Such a powerful effect reflects the importance of strong social ties in scientific activities. First, from a social network perspective, authors embedded in the same institutional social network have strong ties, and intra-institutional academic collaboration and exchange have natural stability and low cost. The citation preference observed in academic papers is an explicit manifestation of informal communication among authors from the same institution. Second, authors working in the same institution typically share similar disciplinary backgrounds, knowledge structures, institutional cultures, and psychological distances, all of which facilitate citation relationship formation. Overall, the model fitting results strongly support the positive effect of formal organizational homophily, thus supporting Hypothesis 3.

The coefficient for variable Nodematch(AU) is negative but not statistically significant, indicating no significant mutual citation tendency among high-impact authors. This conclusion differs from the "citation club" effect [20] supported in some literature, possibly because the model infers overall average trends for the entire sample, while high-frequency mutual citation phenomena within academic elite circles represent only a portion of the overall disciplinary field. Local characteristics cannot be tested through this model and require specific analysis combining actual disciplinary fields and particular groups. Therefore, the empirical results do not support Hypothesis 4.

The significantly negative coefficient for variable Nodematch(CY) indicates that cross-country heterophily promotes citation relationship formation. Specifically, holding other network conditions constant, the odds of citation relationships forming between documents from different countries are 2.344 (= e^0.852) times those between documents from the same country. The model fitting results show that citations are more likely to occur between documents from different countries, contrary to the research hypothesis. While the model results reject country-level document homophily, descriptive analysis earlier revealed ho-

mophily tendencies in a few countries. For example, Tables 2 and 3 show that the United States and a few other countries have very high citation attractiveness—their documents not only have high internal citation rates but also become citation targets for most other countries with relatively lower disciplinary levels. At the country level, scientific research exhibits significant Matthew effects [21], with a few countries becoming disciplinary cores and authorities. Citations flow from peripheral countries to core countries along the gradient of national influence distribution. Combining both analyses, overall citations show country-level heterophily effects, while documents from a few leading countries exhibit some degree of homophily. Based on this, the empirical results partially support Hypothesis 5.

As a control variable, Nodeicov(CT) shows a significantly positive coefficient. For each unit increase in cited document's academic level, the odds of being cited increase by 1.066 times. The estimation results indicate that higher academic value helps increase new citation relationships, with academic value playing a positive role in citation relationship formation, consistent with the basic understanding of citation value orientation.

Conclusions and Implications

This study examines similarity tendencies between citing and cited documents using Exponential Random Graph Models, focusing on the library and information science field. The findings reveal: (1) Overall, document similarity phenomena are ubiquitous in citation relationships and manifest in multiple forms, with triadic structural convergence, journal homophily, and institutional homophily promoting citation relationships; (2) Further analysis shows that document similarity tendencies also exhibit complexity—for example, while the entire sample shows country-level heterophily characteristics, analyzing only the subsample of discipline-dominant countries reveals country-level homophily tendencies.

The empirical results yield the following implications:

(1) **Social proximity is an important formation mechanism for citation relationships.** The citation preferences exhibited by similar documents represent manifestations of social proximity. From an information search perspective, similar documents have advantages in search and identification opportunity costs. Faced with massive literature and complex academic problems, social proximity at any level provides efficient, low-cost guidance to avoid misjudgment risks and academic deviations. Under constraints of time, energy, and knowledge, authors may be more willing to trust and select documents with closer social distance. The social proximity perspective not only provides a good explanation for document similarity phenomena but also enriches understanding of citation formation mechanisms.

(2) **Document similarity accompanying citations results from the interplay between individual intellectual creation and collective**

social factors. Citations are essentially specific relationships resulting from the interaction of multiple elements, such as idea inheritance and collision, academic norm constraints, author influence demonstration, extension of strong ties within organizations, knowledge spillover effects of geographical space, and language-cultural compatibility and inertia. The combined influence of these group-level elements is not only non-negligible but may even exceed usual expectations, as exemplified by the significant and strong effect of institutional homophily. Although conclusions are based on a specific disciplinary sample, they fully reflect that academic creation is not only a mental activity of logical and idea interaction but also a comprehensive result of the social network environment in which authors are embedded, with the influence of network strong ties being particularly important. Both normative theory representing sociology of science and social construction theory are needed for interpretation.

(3) **A network embeddedness perspective may better reveal the complexity of citation behavior.** Citation behavior has complex motivations, and the correlations among citation relationships constitute part of this complexity, which is often neglected in most literature. The emergence of structural embeddedness characteristics of citation relationships in networks reflects interactive influences among author groups, and empirical research fully demonstrates the tendency for citation relationships to embed in triadic structures. Social network and complex network theories provide guidance for network embeddedness analysis, and with the increasing maturity of big data technologies, network modeling based on big data will be an effective approach for future in-depth revelation of citation nature.

Therefore, citation metrics measure not only academic value but more accurately the comprehensive influence of documents—a result of multiple implicit factors. It is necessary to treat citation indicators with objective caution regarding their applicability and interpretation. Additionally, it should be noted that this study focuses on social-level document similarity without addressing topical or content similarity. Future research could expand the analytical framework to incorporate both dimensions of similarity, exploring the interaction between social construction and academic norms in knowledge flow processes. Subsequent studies should also extend empirical validation to other disciplinary fields to test the reliability of research conclusions.

References

[1] Rousso, V., & Quan, W. (2018). Journal impact factor, San Francisco Declaration, and Leiden Manifesto: Commentary and implications. *Library and Information Work*, 62(4), 97-106.

[2] Wang, F. (2016). Core authors in scientometrics from the perspective of integrated publication and citation analysis. *Library and Information Knowledge*,

(1), 4-14.

[3] Colizza, V., Flammini, A., Serrano, M. A., et al. (2006). Detecting rich-club ordering in complex networks. *Nature Physics*, 2(3), 110-115.

[4] Jin, T. (2016). Misuse of self-citation rates in academic journals and countermeasures. *Science and Technology Publishing*, (11), 96-98.

[5] Zhivotovsky, L. A., & Krutovsky, K. V. (2008). Self-citation can inflate h-index. *Scientometrics*, 77(2), 373-375.

[6] Glänzel, W., Debackere, K., Thijs, B., et al. (2006). A concise review on the role of author self-citations in information science, bibliometrics and science policy. *Scientometrics*, 67(2), 263-277.

[7] Bookstein, A., & Yitzhaki, M. (1999). Own-language preference: A new measure of "relative language self-citation". *Scientometrics*, 46(2), 337-348.

[8] Ren, S., & Rousseau, R. (2002). International visibility of Chinese scientific journals. *Scientometrics*, 53(3), 389-405.

[9] Tang, L., Phillips, P., & Youtie, J. (2016). Does China's citation growth exhibit a "club effect"? *Journal of Finance and Economics*, 42(10), 94-107.

[10] Ma, F., & Wu, Y. (2009). A questionnaire survey on paper citation motivations: Taking Chinese journal research and information science communities as examples. *Journal of Intelligence*, (6), 9-14.

[11] Robins, G., Snijders, T., Wang, P., et al. (2007). Recent developments in exponential random graph (p*) models for social networks.* Social Networks*, 29(2), 192-215.

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

[13] Cimenler, O., Reeves, K. A., & Skvoretz, J. (2015). An evaluation of collaborative research in a college of engineering. *Journal of Informetrics*, 9(3), 577-590.

[14] Bornmann, L., & Daniel, H. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1), 45-80.

[15] Rousseau, R., & Liu, Y. (2013). Interestingness and the essence of citation. *Journal of Documentation*, 69(4), 580-589.

[16] Wang, X., Yang, M., Wang, N., et al. (2017). Research project development dynamics of US iSchools under the "Internet Plus" environment. *Information Science*, 35(3), 157-163.

[17] Xiang, X., & Pei, Y. (2016). The impact of geographical proximity on transnational patent collaboration: The mediating role of social proximity. *Science of Science and Management of S.&T.*, 37(4), 17-24.

[18] Snijders, T. A. B., Pattison, P. E., Robins, G. L., et al. (2006). New specifications for exponential random graph models. *Sociological Methodology*, 36(1), 99-153.

[19] Qiu, J., Chen, X., & He, W. (2015). Research on researchers' paper citation motivations and mutual influence relationships. *Library and Information Work*, 59(9), 36-44.

[20] Opsahl, T., Colizza, V., Panzarasa, P., et al. (2008). Prominence and control: The weighted rich-club effect. *Physical Review Letters*, 101(16), 168702.

[21] Yang, X., Gu, X., Wang, Y., et al. (2015). The Matthew effect in China's science: Evidence from academicians of Chinese Academy of Sciences. *Scientometrics*, 102(3), 2089-2105.

Author Contributions

Duan Qingfeng: Research design, data analysis, empirical analysis, and paper writing; Pan Xiaohuan: Data collection and processing.

Empirical Research on the Impact of Document Similarity on Scientific Citation Preference

Duan Qingfeng[1], Pan Xiaohuan[2]

[1] School of Management Science and Engineering, Shanxi University of Finance and Economics, Taiyuan 030006
[2] School of Economics and Management, North University of China, Taiyuan 030051

Abstract: [Purpose/Significance] Understanding the mechanisms behind the similarity phenomenon between citing and cited documents contributes to deeper comprehension of citation nature. [Method/Process] Using Exponential Random Graph Models (ERGM), this study empirically examines the library and information science field to reveal how document similarity affects citation relationships. [Results/Conclusions] Findings demonstrate significant similarity tendencies among citing documents at network structure, institutional, and journal levels. Specifically, citations preferentially embed in transitive triadic structures; documents from identical institutions and journals show higher citation probabilities; and documents from discipline-dominant countries exhibit stronger mutual citation tendencies. These results confirm social proximity as a crucial mechanism in citation behavior formation, reflecting the social attributes of citation preferences.

Keywords: documents; similarity; scientific citation; Exponential Random Graph Model

---

Excellent Reviewers of *Library and Information Work* in 2017

In 2017, over 300 external reviewers participated in the peer review process for *Library and Information Work*, evaluating more than 2,000 manuscripts. Among them, 155 reviewers assessed four or more papers, with an average review time of 6 days. This efficient, high-quality review process ensured the selection of high-quality manuscripts for the journal. Considering review quantity, quality, and timeliness throughout the year, 65 excellent reviewers were selected (listed below). *Library and Information Work* will award certificates to these excellent reviewers and provide free journal subscriptions for one year. We sincerely thank all reviewers for their strong support!

(Listed alphabetically by surname pinyin):

An Xiaomi, Cao Jindan, Chang Chun, Chu Jiewang, Deng Shengli, Ding Yi, Fan Aihong, Gan Chunmei, Gao Fan, Guo Chunxia, Guo Yu, Han Yi, He Sheng, Hu Changping, Hu Zhengyin, Huang Guobin, Teng Guangqing, Wang Cuiping, Song Ge, Wang Jianfang, Sheng Xiaoping, Wu Jianhua, Wu Zhenxin, Xie Rong, Wu Yishan, Wang Lixue, Su Xinning, Wu Zhirong, Qin Hong, Huang ?, Huang Linghe, Xu Haiyun, Xu Xin, Jiang Chunlin, Yan Hui, Li Gang, Li Guojun, Li Jing, Li Jing (Anhui University), Li Rui, Li Wu, Li Yuelin, Liu Bing, Liu Chunli, Liu Hua, Liu Jianzhun, Liu Kan, Liu Xiaojuan, Liu Yu, Liu Yuxian, Liu Ziheng, Mou Dongmei, Pei Lei, Qi Yujie, Qiao Jing, Qin Hong, Shen Lei, Sheng Xiaoping, Song Ge, Su Xinning, Teng Guangqing, Wang Cuiping, Wang Jianfang, Wang Lixue, Wang Yishan, Wu Jianhua, Wu Yishan, Wu Zhenxin, Wu Zhirong, Xie Rong, Yan Hui, Yang Jianlin, Yang Siluo, Yu Liping, Yuan Shunbo, Zhan Qingdong, Zhang Guangqin, Zhao Fei, Zhao Yuxiang, Zheng Dejun, Zheng Qiaoying, Zhou Chunlei, Zhou Qingshan

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