World Journal of Economics and Business Research
Print ISSN: 2960-0073
Online ISSN: 2960-0081
DOI: https://doi.org/10.61784/wjebr3121

INNOVATION MEASUREMENT BASED ON CITATION STRUCTURE AND SYSTEM PERFORMANCE DRIVEN PATH MODELING METHOD

Hao Ma*, RuoNan Wang
School of Business, University of Sanya, Sanya 572000, Hainan, China.
Corresponding Author: Hao Ma

Abstract: To bridge the gap in quantitative innovation analysis and causal pathway modeling for complex citation networks, this study proposes the INFORM framework (integrating Structural Innovation Measurement [OI], Nested Causal Modeling, and TFP Mediation Mechanism). Using USPTO/PATSTAT patent data (426,000 nodes and 5.832 million edges), we empirically verify the "structural innovation → system efficiency → system performance" path through two-way fixed effects and 2SLS methods. Results demonstrate that structural innovation has a statistically significant positive impact on performance ($p<0.01$), with TFP mediation explaining 28-34% of the variance. Four robustness tests confirm the framework's validity. This innovative approach overcomes traditional citation analysis limitations, extends to multi-citation scenarios, and provides a novel tool for patent value assessment.
Keywords: Causal modeling; Mediation mechanism; System performance; Total factor efficiency

1 INTRODUCTION

1.1 Core Issues

The nonlinear structural effect of information diffusion in complex network nodes is the key to system function, but there are two gaps in current research —— the evaluation of node importance lacks innovative value characterization, and the modeling of node indirect interaction lacks path cumulative effect, which affects the mining of core technology patents and the improvement of knowledge graph dissemination efficiency [1].

1.2 Solutions

The INFORM General Modeling Framework is proposed, integrating three core subsystems: citation network structure representation learning, causal path modeling, and efficiency mediation mechanism modeling. Based on node citation structures and citation strength, this framework defines an interpretable structural innovation measurement method. Furthermore, it analyzes the transmission mechanism of this framework to system-level performance indicators using a nested panel regression model [2].

1.3 Three Major Contributions

Develop structural innovation quantification metrics by integrating citation network topology, vector density analysis, and upstream/downstream impact factors to preserve semantic information flow while eliminating noise and statistical redundancy. Construct a multi-layer nested causal modeling framework that establishes a node-to-global performance inference chain: "innovation influence → efficiency variables → system output". Validate pathway effectiveness using large-scale patent citation data, quantify indirect and total effects, and reveal the latent driving force of "original innovation capacity" in enhancing system adaptability.

2 RELATED WORK

In the citation network research, node influence [3], innovation measurement and the correlation between causal path and system performance are three core directions. Existing methods have their own characteristics and limitations, and there are obvious research gaps, as follows:

2.1 Two Core Methods of Node Influence Measurement

The influence of nodes can be categorized into two types: topology-driven and semantic-integrated approaches. The topology-driven approach, represented by metrics like PageRank, HITS, and degree centrality, evaluates node importance based on citation network strength/quantity. While fast in computation and suitable for large-scale networks, it fails to distinguish originality and overlooks citation quality. The semantic-integrated approach, centered around GCN, GAT, and LDA, combines node semantics with topological features to enhance precision and identify semantic differences. However, it suffers from poor interpretability, weak cross-domain transfer capabilities, and high computational costs.

2.2 Evolution of Innovative Measurement Approaches

Innovation measurement encompasses three approaches: text analysis, expert evaluation, and citation structure. Text analysis quantifies new term combinations to objectively identify emerging innovation nodes, though it relies on text quality and struggles to distinguish between "formal novelty" and "substantive innovation." Expert evaluation, typically using the Delphi method, comprehensively identifies implicit innovations across multiple dimensions but suffers from significant subjectivity and scalability limitations. Citation structure measures innovation through metrics like citation frequency, offering accessible data with cross-domain applicability, yet risks conflating "influence" with "innovation" and failing to establish causal relationships with systemic performance.

2.3 Existing Deficiencies in Causal Path and System Performance Correlation

Existing research exhibits three critical limitations: First, traditional causal mediation models focus solely on macro-level aggregate relationships while neglecting structural innovation at the node level, failing to account for micro-level influences on system performance. Second, their poor adaptation to high-dimensional citation networks makes it difficult to quantify causal effects of structural innovations and verify the complete path from "innovation → mediating variables → system performance". Third, they only detect correlations between innovation and system performance without distinguishing between causal relationships and spurious associations, resulting in insufficient causal explanations.

2.4 Core Research Gaps

Existing research faces three major gaps: First, there is a lack of systematic measurement for "structural innovation" that defines innovation through the perspective of citation network topology transformation. Second, no universal causal framework has been established linking "structural innovation → mediating variables → system performance", leaving the transmission mechanisms of mediating variables unclear. Third, the integration between graph structure modeling and causal inference remains inadequate [4], failing to balance explanatory power with adaptability to citation networks.

3 METHODOLOGY

The INFORM framework proposed in this study aims to quantify the structural innovation of large-scale citation networks and evaluate the causal transmission of system performance through the mediation of efficiency [5]. The core contains two modules: structural innovation measurement and nested causal modeling.

3.1 Structural Innovation Measurement

Core objectives and overall composition of the framework: INFORM framework aims to quantify the structural innovation of large-scale citation networks and evaluate the causal transmission of system performance through efficiency mediation. It consists of two modules: "structural innovation measurement" and "nested causal modeling framework".

Details of structural innovation measurement (OI): Data anchoring: 2000-2022.6 A share patent data and CNRDS specifications;

Four key modeling principles: Citation density factor (quantifying upstream-downstream redundant correlations with minimal redundancy for stronger breakthrough capability), citation path depth (using June 2022 as the temporal window to identify maximum citation tiers while excluding outliers), normalized citation intensity (logarithmic standardization of citation frequency adjusted for observation duration to eliminate temporal bias), and OI synthesis (logarithmic synthesis using three principal component weights extracted through PCA).

Advantages and uses of OI: It is better than the traditional citation measure, with both topological expression and temporal adaptability. As a core explanatory variable, it verifies the driving effect on the financial performance of listed companies (such as lnroe).

3.2 Details of the Nested Causal Modeling Framework

Sample anchoring: 2000.12022.6 A-share listed companies sample;

Three-level modeling framework: Level 1 (Direct Effects): Without introducing mediating variables, analyze the direct impact of OI on system outputs (logistic nonlinear returns on input, operational efficiency, etc.) while controlling for variables such as scale and leverage ratio. Level 2 (Mediation Modeling): Using Total Factor Productivity (TFP) as a mediator to decompose "OI → TFP" (forward chain, verifying innovation-driven efficiency) and "TFP → output" (backward chain, verifying efficiency-to-performance transmission). Level 3 (Nested Path Identification): Combine direct and mediated effects by quantifying their proportions through stepwise regression and Sobel test. Model specification and variable control: Employ bidirectional fixed effects with clustered robust standard errors to address heteroskedasticity/serial correlation. Introduce 10 control variables including innovation investment and equity concentration to avoid omitted variable bias.

3.3 Construction of Efficiency-driven Transmission Path (Efficiency Driven Transmission Path)

System efficiency serves as a pivotal mediating variable bridging structural innovation and organizational performance [6]. In information systems, Total Factor Productivity (TFP) is widely used to characterize the output capacity achievable per unit of resource input. Fluctuations in TFP reflect both the quality of resource allocation and the extent of process optimization.

This study uses total factor efficiency based on input-output system perspective as the mediating mechanism, mainly considering the following points:

1) System efficiency can be affected by structural innovation: high structural innovation nodes are often accompanied by better knowledge combination paths, triggering the optimization of resource allocation efficiency;

2) System efficiency further affects the final output: efficient use of resources will improve the output per unit time, response speed and overall performance;

3) The intermediary path has the advantage of interpretation: it helps to reveal the underlying mechanism of "structural behavior → system response"

In order to further improve the identification strength, the model introduces the structural innovation and industry reference average of one period lag as instrumental variables, and uses 2SLS estimation method to alleviate endogeneity bias [7].

Firstly, model 2 is constructed to test the direct impact of total factor productivity on resource allocation efficiency [8]:

$$\text{lnroe} = \theta_0 + \alpha \text{lnrate} + \sum \text{CV} + \varepsilon_{it} \tag{1}$$

Secondly, the direct impact of TFP as a proxy variable for resource allocation efficiency is adopted in model 3:

$$\text{lntfp\_lp} = \theta_0 + \beta \text{lnrate} + \sum \text{CV} + \varepsilon_{it} \tag{2}$$

Finally, the intermediate variable of system efficiency is included in model 1 to test whether there is a systematic efficiency mediation effect, and model 4 is established:

$$\text{lnroe} = \theta_0 + \gamma \text{lnrate} + \sum \text{CV} + \text{lntfp\_lp} + \varepsilon_{it} \tag{3}$$

In the above model, subscript i and t represent city and γtime; α, β, are the regression coefficients of each variable; CV is the control variable; θ is the random disturbance term.

Table 1 Test of the Mediating Effect of Systematic Efficiency on Financial Performance: Stepwise Method

class	**model** (1)	**model** (2)	**model** (3)
	lnroe	**lntfp_lp**	lnroe
atelnr	0.018* (0.007)	0.020* (0.007)	
lntfp_lp			0.018* (0.007)
CV		YES	

Note: *P <0.1, P <0.05, *P <0.01; the value in parentheses is the standard error of regression coefficient.

The empirical results in Table 1 demonstrate that Model (2), which uses Total Factor Productivity (TFP) as a proxy for resource allocation efficiency, shows a positive regression coefficient of 0.020 at the 1% significance level. This indicates that improved resource allocation efficiency enhances total factor productivity (TFP), supporting Hypothesis H2. This effect may stem from innovations in technologies and processes that optimize resource distribution. Model (3) examines the mediating role of total factor productivity's systematic efficiency, revealing a significant positive coefficient of 0.018 at the 1% level. This suggests that TFP improvement significantly boosts resource allocation efficiency, validating Hypothesis H3. The mechanism likely occurs through cost reduction and quality enhancement, which translate into better financial performance. Combining the findings from Models (1)-(3), we observe that a 1% increase in TFP leads to a 0.018% improvement in resource allocation efficiency, further confirming Hypothesis H1.

3.4 Control Variable Setting and Robustness Design

3.4.1 Setting of control variables

To ensure the accuracy and reliability of model estimation, this study incorporates multi-dimensional control variables: structural scale (system size, number of nodes), resource allocation (resource utilization ratio, structural leverage), governance mechanisms (concentration level, upstream-downstream control chain length), and external conditions (citation network density, temporal window position). Four robustness tests are employed: replacing endogenous variables (substituting with economically equivalent indicators), excluding exceptional period samples (excluding extreme impact years), instrumental variable regression (to mitigate endogeneity), and mediation path validation (repeated verification of core transmission paths).

3.4.2 Robustness tests: variable substitution, elimination of abnormal periods and instrumental variable estimation

To ensure robust conclusions, this study conducted robustness tests through indicator substitution, period-specific sample exclusion, and instrumental variable estimation. Specifically, core explanatory variables (e.g., Indicator A) were replaced with alternatives sharing similar economic implications (e.g., Indicator B). As shown in Table 2, the coefficient of the core independent variable remained statistically significant at the 1% level (0.012) after variable transformation, demonstrating that the findings remain unaffected by measurement methods and exhibit high robustness.

Table 2 Results of Robustness Test

Class	lnroa		
	model 1	model 2	model 3
	Excluding macro influencing factors I	Excluding macro factors II	I+II
lnrate	0.023***	0.021***	0.027***
	(0.007)	(0.008）	(0.008）
lnrd	-0.020***	-0.013***	-0.016***
	(0.003)	(0.004）	(0.003)
lnscale	0.006**	0.007**	0.010***
	(0.003)	(0.003）	(0.003)
llnleverage	0.103***	0.134***	0.098***
	(0.015)	(0.017）	(0.002)
lnhold	0.011***	0.009***	0.010***
	(0.001）	(0.002）	(0.001)
lnocf	-0.019***	-0.017***	-0.017***
	(0.0009）	(0.0009）	(0.001)
lnta	-0.012***	-0.019***	-0.018***
	(0.003）	(0.003）	(0.003)
llnconcent	0.022***	0.031***	0.028***
	(0.006）	(0.006）	(0.006)
Relate	-0.004	-0.002	-0.006
	(0.004）	(0.003）	(0.004)
Dualjob	0.0009	0.002	0
	(0.002）	(0.002）	(0）
Quality	0.005	0.003	0.010
	(0.005）	(0.006）	(0.007)
_cons	0.226***	0.278***	0.282***
	(0.062）	(0.066）	(0.067)
Urban fixed effects	YES	YES	YES
Time fixed effects	YES	YES	YES
N	14758	11377	9827

Note: *P <0.1, **P <0.05, ***P <0.01; the value in parentheses is the standard error of regression coefficient.

The INFORM framework integrates structural innovation metrics, efficiency mediation mechanisms [7], and causal path decomposition into a unified system that enables generalization, interpretability, and empirical verification. This methodology not only applies to citation networks but can also be extended to behavioral propagation modeling tasks in other temporal graph systems.

4 EXPERIMENTAL DESIGN AND RESULT ANALYSIS

This part focuses on the experimental verification of the proposed INFORM framework. Using large-scale citation network data, multiple regression models are constructed to systematically test the causal transmission path of "structural innovation → system efficiency → system performance", and carry out robustness and heterogeneity analysis.

4.1 Experimental Environment (Experimental Environment)

4.1.1 Hardware platform

Server CPU: Intel Xeon Gold 6248R (24 cores, 48 threads, base frequency 3.0GHz)
Memory: 128GB DDR4 ECC memory (2666MHz)
Graphics card: NVIDIA Tesla V100 (32GB HBM2 memory for network structure feature calculation acceleration)
Storage: 2TB SSD (for large-scale patent data storage and fast read)

4.1.2 Software tools

Programming languages and libraries: Python 3.8 (NetworkX 2.6.3 for network construction, PyTorch 1.10.0 for feature extraction, Pandas 1.4.2 for data preprocessing)
Statistical modeling tools: Stata 17 (two-way fixed effects regression, 2SLS instrumental variable estimation, mediation effect test)
Efficiency calculation tool: DEASolver Pro 5.0 (assists in the calculation and verification of total factor productivity (TFP))

4.2 Data Set (Dataset)

4.2.1 Data sources

The experiment uses two public patent citation databases:

USPTO Patent Citation Database (2010-2020): covers the citation relationship and technical attributes of authorized patents in the United States;

PATSTAT Global Patent Database (2010-2020): Supplementing patent data from Europe, China, and other regions to ensure cross-regional representativeness.

Time span: 2010-2020, using a three-year rolling time window to divide the dynamic panel (2010-2012,2013-2015,2016-2018,2019-2020, a total of 4 periods).

4.2.2 Data characteristics and cleaning

Node definition: A single patent entity is a node, including patent application number, application date, technical classification number and other attributes;

Edge definition: If patent A cites patent B, then construct a directed edge from B to A (representing the direction of knowledge dissemination);

Cleaning strategy: ① Remove isolated nodes (patents with no in/boundary edges); ② Remove extreme values (patents whose citation frequency is> 99% or <1% of the percentile); ③ Remove missing information samples (patents with missing key attributes such as application number and technical classification number);

Final scale: the number of nodes is 426,000, and the number of citations is 5.832 million, forming a time-oriented citation network.

4.3 Experimental Process (Experimental Process)

4.3.1 Data sources and preprocessing

Data base: based on the USPTO/PATSTAT published patent citation database, covering more than 10 years in multiple fields, ensuring the representativeness and time continuity of the sample.

Data features: nodes are patents, edges are citation relationships; a three-year rolling time window is set to transform dynamic panel data, and isolated nodes are removed to reduce noise.

Data scale: After preprocessing, a high-quality network with hundreds of thousands of patent nodes and millions of citation edges is formed, which takes into account both representativeness and scalability.

4.3.2 Variable design and index construction

Core explanatory variable (structural innovation OI): Based on the measurement logic in Section 3, it integrates three-dimensional features such as citation density factor, propagation depth and time window weight to capture the breakthrough of patents in the citation topology.

Intermediate variables (system efficiency TFP): measured by Levinsohn-Petrin method, the input is the number of patent applications and technical classification numbers, and the output is the frequency of patent citation and depth of citation. It is embedded in the mediation regression in panel form.

The dependent variable (performance) includes four dimensions: technology influence (growth rate of patent citation), network centrality (PageRank score), node reach, and subsequent citation success rate.

Control variables: node level (patent active time, type, IPC classification level), network level (citation network density), bidirectional fixed effects of technology field × year and time dummy variable.

4.3.3 Model specification and estimation method

Core model system: construct four progressive models to test the transmission path —— direct effect model (OI→system performance), mediation model I (OI→TFP), mediation model II (TFP→system performance), and complete mediation model (OI→TFP→system performance).

Selection of estimation methods:

Two-way fixed effect model (Two-way FE): control the constant factors of technology field and year;

Instrumental variable estimation (2SLS): using the lag of innovation and the average value of the field as instrumental variables to alleviate endogeneity;

Bootstrap effect decomposition (Bootstrap + Sobel test): identify direct and indirect effects.

4.4 Analysis of Empirical Results

4.4.1 Principal regression results

Structural innovation has a significant positive impact on system performance ($p < 0.01$);

Structural innovation has a strong positive explanatory force on efficiency variables, which verifies that it is a potential driving factor;

System efficiency significantly affects system performance and plays a mediating role in path analysis.

The proportion of intermediary effect is between 28% and 34%, indicating that structural innovation not only affects performance itself, but also produces indirect benefits by improving the efficiency of system resource allocation.

4.4.2 Robustness tests

The following strategies are used to ensure robustness of estimates:

Replace performance variables (e.g. ROA → Cascade Size);

Delete samples of abnormal years (such as epidemic and economic crisis years);

Replacement efficiency calculation method (such as DEA → LP method);

Introduce lagged variable/instrumental variable regression to eliminate endogeneity bias.
All regression directions are consistent, the significance level is stable, and the robustness of the verification model is excellent.

5 CONCLUSION AND FUTURE WORK

5.1 Research Conclusions and Core Contributions

The core contributions and conclusions of this study are as follows: First, a multi-dimensional structural innovation measurement method is proposed to integrate citation topology, temporal path and propagation breadth, so as to break through the limitations of traditional citation frequency and accurately capture the "breakthrough" value of nodes in the citation network;
The second is to construct a nested causal reasoning framework [9], with system efficiency as the mediator, forming a path of "structural innovation→ system efficiency→ system performance", using two-way fixed effects to control interference and instrumental variables to slow endogeneity, so as to accurately identify causal associations;
Thirdly, the empirical verification mechanism demonstrates stable universality. Based on large-scale patent data, it confirms that structural innovation positively drives system performance, with efficiency mediation pathways proving effective across different technical domains and subgraph structures, demonstrating the robustness and generalizability of the INFORM framework. This study fills the methodological gap in "quantifying causal pathway identification for structural innovation," providing a new paradigm for complex system modeling that combines theoretical and practical value.

5.2 Research Limitations and Future Directions

Although the INFORM framework constructed in this paper has a systematic modeling and empirical path, it still needs to be improved. In the future, it can be expanded from four aspects:
Semantic enhancement of citation relationship: At present, only graph structure features are used to measure innovation. In the future, semantic information such as abstract and topic vector will be integrated to build a "structure-semantics" joint embedding model to improve measurement accuracy and index interpretability [10].
Model diagram learning and expansion: The existing panel regression model is inefficient in large-scale graphs. In the future, graph neural network (GNN) will be combined to obtain innovative features through semi-supervised/unsupervised learning, so as to balance interpretability and computational efficiency [11].
Cross-domain expansion of application scenarios: Currently, only patent citation network is verified, and in the future, it will be extended to academic literature, enterprise knowledge graph, technology roadmap and other scenarios to expand application boundaries and explore adaptability.
Deep extension of causal chain: the existing "structural innovation→ efficiency→ performance" mediation model is relatively simple. In the future, multiple mediating paths, moderating variables and dynamic feedback mechanism will be introduced to construct a complex causal map and conform to the real logic of the system.
This study pioneers the introduction of structural innovation metrics in citation networks, clarifying their driving mechanisms for system outputs. The INFORM framework combines theoretical innovation with engineering applicability, serving as a versatile foundation for graph-structured causal modeling with broad prospects. Subsequent research will focus on semantic enhancement, algorithm optimization, and scenario adaptation to deepen exploration and practical applications.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Zhang Meng. Research on Reinforcement Learning Algorithms Based on Causal Models. Beijing University of Posts and Telecommunications, 2025(6). DOI: 10.26969/d.cnki.gbydu.2024.001473.
[2] Zhang Huiguo. Statistical Analysis and Application of Multi-level Nested Panel Data. Xinjiang University, 2006.
[3] Jin Sheng. Research on Key Technologies for Measuring Influence of Complex Network Nodes. Qinghai Normal University, 2025.
[4] Luo Wei, Liu Yu, Huang Qiang, et al. A Map Matching Method Integrating Graph Structure Learning and Lightweight Recurrent Modeling. Peking University Journal (Natural Science Edition), 2024, 60(06): 979-988.
[5] Li Xiao. Research on the Impact of R&D Investment on Corporate Performance. Shandong University, 2024.
[6] Guo Yifan, Tian Jiahui. Research on the Impact of Personalized Recommendation Systems on Consumer Purchase Decision Efficiency. Modern Marketing, 2025(27): 154-156.
[7] Zhang Min, Li Yan. The Impact of the Digital Economy on the Efficiency of Commercial Circulation—and the Mediating Mechanism of Technological Innovation and Management Innovation. Journal of Commercial Economic Research, 2023(18): 19-22.

[8] Ozherlin, Hou Lingling, Yi Yuxi. Research on the Impact of Digital Transformation on Total Factor Productivity in Enterprises—Based on Empirical Evidence from Listed Companies in Hunan Province. Business Economics, 2025(11): 181-184.
[9] Yan Yue-rong, Liu Dong-liang. Limitations and Overcoming of Big Data Analysis in Legal Field Applications. Journal of Xi'an Jiaotong University (Social Sciences Edition), 2025: 1-14.
[10] Ji W L, Yu D H, Zhang W S. Knowledge Graph Embedding Model Based on Dual Relationship Type Constraints. Microelectronics & Computer, 2025, 42(08): 29-37.
[11] Guan H S, Bao Q, Wang Q. Study on Resilience Evaluation and Evolutionary Characteristics of Regional Digital Innovation Ecosystems—Based on Panel Vector Autoregression Model. Statistics and Management, 2025, 40(09): 28-40.