Multi-Scale Temporal Scenario Planning for Financial Networks: A GNN Approach to Stress Testing

Xinyan Zhang[1] **and Xiaobing Feng**[1] **and Xiujuan Xu**[2,*] **and Ling Feng**[3] **and Jinghua Lian**[4]
[1,2] Dalian University of Technology
[1,3] Shanghai University Of International Business Economics
[4] Guangzhou University
[1]dahuzidahu@gmail.com,[2]xjxu@dlut.edu.cn [1] fxb@suibe.edu.cn,
[3]13120925823@163.com,[4]lianjinghua@e.gzhu.edu.cn

Abstract

Financial networks have grown increasingly complex and interconnected, creating urgent challenges for systemic risk management. We propose a robust multi-scenario stress testing framework based on graph neural networks that enables large-scale anomaly detection and systematic risk assessment across pre- and post-pandemic financial landscapes. Our approach integrates several technical innovations: efficient sparse matrix computation for graphs with over 81,434 nodes, dynamic class imbalance handling that improves recall by nearly 17 times, and a comprehensive scenario-based evaluation protocol examining baseline performance, feature noise resilience, structural vulnerability, and susceptibility to information shocks. Experiments on real financial data comparing the 2019 (pre-pandemic) and 2022 (post-pandemic) periods reveal a significant shift in risk characteristics – post-pandemic networks demonstrate heightened vulnerability to structural changes (-9.4% AUC-PR) and information propagation (-3.9% AUC-PR), indicating that risk sources have evolved from data quality concerns to network connectivity and information flow dynamics. Our framework provides regulators and financial institutions with practical tools to identify emergent risks and enhance system resilience against future structural and information-based shocks.

Keywords: GNN, multi-scale scenario planning, fake news detection in finance

1 Introduction

The increasing complexity and interconnectedness of global financial systems have made stress testing a crucial tool for identifying systemic risks and supporting macroprudential policy(Pritsker, 2011; Federal Reserve System, 2024; European Banking Authority, 2016, 2024; Bank of England, 2022). Traditional stress testing frameworks, however, often rely on macroeconomic variables and static scenario design, limiting their ability to address heterogeneous, technology-driven, or structural risks. Recent events such as the COVID-19 pandemic have further highlighted the need for dynamic, multi-scenario approaches that can capture evolving risk transmission paths and the impact of information shocks(Lim, 2016; Bank of Japan, 2024).

To address these challenges, we develop a graph neural network (GNN)-based framework for large-scale financial anomaly detection and multi-scenario stress testing. Our method features: (1) scalable processing of financial graphs with over 80,000 nodes and 350,000 records via sparse matrix and memory optimization; (2) a dual-weighting mechanism combining dynamic class weights and improved Focal Loss to tackle severe class imbalance; (3) an adaptive threshold selection algorithm to optimize precision-recall trade-offs; and (4) a scenario design covering baseline, feature noise, graph structure change, and fake news propagation, enabling systematic evaluation of network vulnerability and resilience.

We fuse multi-source data (structured financials, sentiment, regulatory records) and employ PCA for efficient feature engineering, retaining over 91% of information. Comparative experiments on pre- and post-pandemic data (2019 vs. 2022) show that post-pandemic financial networks exhibit much higher sensitivity to structural and information shocks (AUC-PR drops of -9.4% and -3.9%, respectively), indicating a shift in risk sources from data quality to network connectivity and information flow. These findings suggest the need for enhanced monitoring of network structure and information propagation in financial regulation.

Our contributions are threefold: (1) a scalable GNN-based anomaly detection and stress testing framework for large financial networks; (2) methodological innovations in class imbalance handling and scenario-based evaluation; (3) empirical evidence of evolving risk characteristics in financial

Proceedings of the 2nd Agent AI for Scenario Planning (AgentScen), Montreal, Canada, August 16, 2025

systems under systemic shocks. Model capacity's impact on performance is summarized in Appendix Table A.4. The proposed approach offers both technical solutions and policy insights for improving financial system resilience.

2 Related Work

In recent years, literature on stress testing has evolved toward integrating agent-based modeling (ABM) and graph neural networks (GNNs) to address complex systemic risks. For example, Samimi et al. (2024) demonstrated how Agent-Based Modeling (ABM) can simulate autonomous agent behaviors and interactions to enhance system safety and risk management, while Bernárdez et al. (2023) proposed MAGNNETO, a distributed GNN-multi-agent framework for traffic engineering optimization. These studies highlight the potential of hybrid models that combine agent autonomy with graph-based structure learning.

2.1 Classical Theory and Basic Definitions

Pritsker (2011) is an important representative figure in stress testing theory construction. His proposed "Enhanced Stress Testing" framework emphasizes a risk exposure-driven system modeling approach, distinct from traditional linear models that rely solely on macroeconomic variable shocks. His research particularly proposed the concept of "Trust Set," which involves constructing a set of reasonable but non-unique scenarios to conduct multi-dimensional shock resistance assessments of institutions under highly uncertain environments, enhancing the robustness of testing.

2.2 U.S. Stress Testing System Experience

The Federal Reserve System has established a comprehensive modeling framework encompassing modules for loan and trading losses, net income, and capital adequacy. This approach emphasizes scenario design based on historically extreme but plausible events, data-driven modeling, and institutional independence, while employing unified tools to assess multi-institutional responses and balancing regulatory transparency with market stability (Federal Reserve System, 2024). Furthermore, regulatory provisions "Rules and Regulations (6651–6664)" (Federal Register, 2023) highlight public participation, model updates, and risk evolution, underscoring the normative and progressive features of the U.S. system.

2.3 Comparison of EU and UK Approaches

The European Banking Authority's "2025 EU-wide Stress Test Methodological Note" advocates incorporating structural shocks, such as climate change, into stress testing and emphasizes consistent cross-national assessment (European Banking Authority, 2024). The earlier "2016 FAQ document" established procedures for identifying capital adequacy, risk concentration, and contagion paths, laying the foundation for institutionalized stress testing (European Banking Authority, 2016).

The Bank of England, in its "2022 Annual Cyclical Scenario (ACS) Elements Description," highlights the evaluation of structural and non-linear risks through multi-path carbon policy simulations and adaptive balance sheet assessments, exemplifying climate stress testing practices (Bank of England, 2022).

2.4 Emerging Explorations in Asia

The Monetary Authority of Singapore has expanded stress testing to include technological risks such as AI model errors and cyber attacks, demonstrating forward-looking regulatory awareness (Lim, 2016). The Bank of Japan's 2024 "Financial System Report" analyzes the long-term effects of population aging on the financial system, highlighting structural risks to bank capital adequacy and adaptation strategies for financial institutions (Bank of Japan, 2024).

2.5 Other Methodological Extensions

At the investment management level, Ruban and Melas (2010) proposed using multi-factor risk models to conduct stress assessments of investment portfolios, emphasizing risk factor linkage mechanisms and the adaptability of micro-asset allocation, which is an important complementary path for micro-financial stress testing (Ruban and Melas, 2010).

Ok and Eniola (2025) proposed a deep learning-based scenario reasoning method in their research, using unstructured data to enhance the model's sensitivity and response capability to emerging risks, demonstrating the potential of AI tools in cross-variable modeling and data dimension adaptation (Ok and Eniola, 2025).

In the field of graph neural networks and financial risk detection, Weber et al. (2019)first applied GCN to financial network analysis, demonstrating the effectiveness of graph structure information in capturing financial anomalies, although their research was limited to small-scale data. Thilagavathi et al. (2024) proposed a framework combining graph neural networks and anomaly detection techniques for financial fraud detection, achieving a 95% detection rate on highly imbalanced credit card fraud datasets, but mainly focused on credit card transactions without extension to more complex financial network structures. Balmaseda et al. (2023) explored the application of deep graph learning in predicting systemic risks in financial systems, emphasizing the importance of machine learning in analyzing large financial networks, but traditional techniques still have limitations in handling complex relationships. While these studies have advanced the application of graph neural networks in the financial domain, they still showed obvious deficiencies in processing large-scale data, solving extreme class imbalance, and constructing multi-scenario stress testing frameworks.

Finally, in the field of behavioral finance and psychology, Ward et al. (2021) discussed behavioral response mechanisms under system shocks in their chapter, emphasizing the important influence of institutional resilience and psychological coping abilities on stress test assessment results, providing an important literature foundation for expanding the social dimension of stress testing.

3 Methodology

3.1 Overall Research Framework

This research proposes a large-scale financial anomaly detection and stress testing framework based on graph neural networks, mainly divided into two core tasks: 1) large-scale financial graph anomaly detection; and 2) multi-year, multi-scenario financial system stress testing. The overall framework proceeds in three stages: data processing, model construction, and result evaluation.

The research framework first preprocesses the original financial data, including data cleaning, feature engineering, and graph structure construction, then designs corresponding graph neural network models and optimization strategies for the two main tasks, and finally evaluates model performance through comprehensive evaluation metrics.

The two core tasks have different focuses: Task 1 focuses on the micro-level identification of anomalous entities, addressing challenges such as large-scale financial graph data processing, extreme class imbalance, and recall improvement; Task 2 takes a macroprudential perspective, evaluating the vulnerability and resilience of financial networks in different periods through the construction of a multi-scenario stress testing framework.

3.2 Data Preprocessing and Feature Engineering

3.2.1 Data Cleaning

Original financial data typically contains noise, missing values, and outliers that require cleaning. In this study, missing values (NaN), positive infinity, and negative infinity were replaced with 0.0 to ensure data completeness. Outliers were handled by standardizing all features to have a mean of 0 and variance of 1, reducing their influence and making features comparable.

3.2.2 Feature Engineering

This study used two main feature dimensionality reduction methods:

Principal Component Analysis (PCA): Through linear transformation, the original high-dimensional features (28 dimensions) were reduced to 15 dimensions while retaining approximately 91.37% of the information. PCA preserves the principal components that maximize data variance, helping to reduce feature redundancy and improve computational efficiency.

Nonlinear kernel dimensionality reduction (Nyström method): This method first uses the Nyström algorithm to approximate the RBF kernel function mapping to high-dimensional space and then applies PCA dimensionality reduction, which can better capture nonlinear relationships between features. This method effectively reduced computational complexity while maintaining approximately 85.59% of the original information.

A comparison of the two methods found that linear PCA not only retained a higher proportion of data variance but also had high computational efficiency and strong interpretability of principal components, so PCA dimensionality reduction was mainly used in subsequent experiments.

3.2.3 Graph Structure Construction

This study constructed graph structure networks through common behaviors between users (such as following the same stocks). Specifically, if two users followed the same stock, a connection relationship was established between them. This construction method is based on the assumption that users who follow the same stocks may have similar behavioral patterns or risk characteristics.

The adjacency matrix was stored in sparse matrix format, with each non-zero element representing a connection between two user nodes. For large-scale datasets (such as Task 1's 350,000 records), this sparse representation method greatly reduced storage and computational overhead.

3.3 Task 1: Large-Scale Financial Graph Anomaly Detection Method

3.3.1 Supervised Graph Neural Network Model

The supervised graph neural network model designed in this study mainly includes three layers of graph convolutional networks (GCN), with LayerNorm standardization between layers, supplemented by residual connections and multi-layer classifiers.

The mathematical representation of the graph convolutional layer is:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{1}$$

Where $\tilde{A} = A + I_N$ is the adjacency matrix with self-loops added, $\tilde{D}$ is the corresponding degree matrix, $H^{(l)}$ is the node feature matrix of the l-th layer, $W^{(l)}$ is the learnable weight matrix, and σ is the nonlinear activation function (ReLU is used in this study).

The main features of the model include: **Three-layer graph convolutional network** capturing high-order graph structure information through multiple layers of convolution, with adjustable output dimensions for each layer (such as 64/96/128/192); **Residual connection** directly connecting the output of the first layer to the output of the third layer, in the form: $H^{(3)} = H^{(3)} + H^{(1)}$, which helps alleviate training difficulties in deep networks and promotes gradient flow; **LayerNorm instead of BatchNorm** used for standardization after each graph convolutional layer, in the form:

$$\text{LayerNorm}(x) = \gamma\frac{x-\mu}{\sqrt{\sigma^2+\epsilon}} + \beta \tag{2}$$

where μ and σ are the mean and standard deviation along the feature dimension, and γ and β are learnable parameters; and **Multi-layer classifier** using a two-layer fully connected network, with the first layer having the hidden dimension and using ReLU activation, and the second layer outputting a single scalar value representing the probability of a node being anomalous.

3.3.2 Imbalanced Sample Handling

To address the severe class imbalance problem in financial anomaly detection (abnormal samples accounting for only 5.29%), this study adopted two main strategies:

Class weighting: Sample weights are dynamically calculated based on the ratio of positive to negative samples, using $\text{weight}_{\text{pos}} = \text{balance\_ratio} \times \frac{n_{\text{neg}}}{n_{\text{pos}}}$ and $\text{weight}_{\text{neg}} = 1.0$, where balance_ratio is an adjustable parameter. In Task 1's dataset, the positive sample weight was approximately 4.88 times that of the negative samples.

Improved Focal Loss: Assigning higher loss weights to hard-to-classify samples (especially minority classes), with the formula:

$$\text{FL}(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t) \tag{3}$$

Where p_t is the predicted probability of a sample belonging to its true class, α_t is the class weight (set to 0.75 in this study, giving more attention to anomalous samples), and γ is a modulation parameter (set to 2.0), controlling the rate at which the weight of easily classified samples decreases.

This dual-weighting mechanism made the model pay more attention to minority class samples during the training process, effectively enhancing the ability to identify anomalous samples.

3.3.3 Large-Scale Graph Data Memory Optimization

To process large-scale financial graph data containing hundreds of thousands of nodes, we implemented several memory optimization strategies. These include sparse adjacency matrix representation, adjacency matrix normalization, regular garbage collection, and full graph training rather than batch training. This approach enabled processing graphs with over 80,000 nodes within reasonable memory constraints. Further details on these optimization techniques are provided in Appendix A.4.

3.3.4 Optimal Threshold Selection Method

To achieve the best classification effect on imbalanced datasets, this study implemented an automatic threshold selection algorithm. This method finds the best decision threshold based on performance on the validation set, rather than using the default 0.5.

For threshold selection based on the F1 score, the algorithm calculates precision and recall under different thresholds, calculates the corresponding

F1 score, and selects the threshold that maximizes the F1 score. For threshold selection based on the G-Mean, the algorithm calculates recall and specificity under different thresholds, calculates the corresponding G-Mean, and selects the threshold that maximizes the G-Mean.

In financial anomaly detection scenarios, this adaptive threshold method can better balance precision and recall, significantly improving the practical utility of the model.

3.4 Task 2: Multi-Year Multi-Scenario Stress Testing Method

3.4.1 Enhanced Graph Neural Network Model

Task 2 added two specially designed components to the basic model of Task 1 to enhance the model's adaptability to different stress scenarios:

Attention mechanism: Introducing attention weights for each node's features, allowing the model to automatically focus on the most important feature dimensions. The attention calculation process is as follows:

$$a_i = \sigma(W_2 \cdot \mathrm{ReLU}(W_1 \cdot h_i)) \quad (4)$$

$$h'_i = h_i \odot a_i \quad (5)$$

where h_i is the feature vector of node i, W_1 and W_2 are learnable weight matrices, σ is the sigmoid activation function, and $\odot$ represents element-wise multiplication.

Fake news filter: A special gating mechanism that learns to suppress features that may be noise or anomalies. The filtering process is:

$$g_i = \sigma(W_4 \cdot \mathrm{ReLU}(W_3 \cdot h_i)) \quad (6)$$

$$h''_i = h'_i \odot g_i \quad (7)$$

where g_i is the filter gate value, and W_3 and W_4 are learnable weight matrices.

These two components used in combination enable the model to focus on the most relevant features and nodes through the attention mechanism and learn to suppress features that may be noise or anomalies through the fake news filter, enhancing the model's adaptability to stress scenarios.

3.4.2 Multi-Scenario Stress Testing Framework

This study's stress testing framework draws on the mainstream scenario planning pipeline concept. Specifically, external shocks (such as the COVID-19 pandemic) first transmit through the global financial market to the local financial system, then affect market entities and their responses to fake news, forming a closed loop of forward and reverse risk transmission. This study designed four typical stress test scenarios to systematically assess the vulnerability and resilience of the financial system: (1) **Baseline scenario:** No external interference, serving as a reference standard; (2) **Feature noise scenario:** Simulating data quality decline or market fluctuations by adding random Gaussian noise with intensity 0.1 to the original features; (3) **Graph structure change scenario:** Simulating financial network connection breakage or institution collapse by randomly removing 20% of the edges; and (4) **Fake news propagation scenario:** Simulating market panic or rumor spread, triggered from a small number of initial nodes (about 1%), with propagation probability 0.7, influence intensity 0.3, simulating the information diffusion process through the network structure.

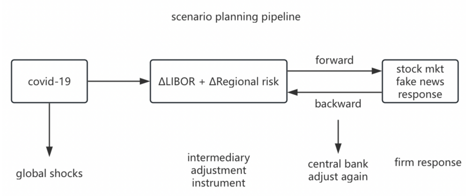

Figure 1: Stress testing scenario planning pipeline. External shocks (e.g., COVID-19) transmit through global financial markets to the local financial system, impacting market entities (e.g., stock market, firms, and responses to fake news), with feedback loops illustrating forward and backward risk transmission.

Additional details on the fake news propagation simulation and temporal comparison analysis methods are provided in Appendix A.5.

3.4.3 Rationale and Justification for Stress Test Scenario Parameters

The selection of appropriate parameters is fundamental to the validity of the stress-testing framework. This section, therefore, provides a detailed justification for the key parameters (feature noise intensity, edge removal rate, and fake news propagation parameters) used in our stress tests. All parameters are chosen to simulate "severe but plausible" conditions, a core principle in financial stability assessment and regulatory stress testing (BPI Staff). Our choices are informed by academic literature, industry practice, and the specific objectives of each scenario.

1. Feature Noise Scenario: Noise Intensity = 0.1 We introduce Gaussian noise with an intensity of 0.1 to the node feature vectors. This choice is motivated by two primary considerations:

- *Simulating moderate data quality issues and market volatility*: Real-world financial data is subject to noise from reporting delays, measurement errors, or short-term irrational sentiment. An intensity of 0.1 represents a moderate disturbance, not catastrophic data corruption, and serves to test the model's *robustness* against common data imperfections. Robustness—the ability to maintain performance under common corruptions or perturbations—is a key aspect of real-world reliability (Hendrycks and Dietterich, 2019).

- *Data augmentation and regularization*: Adding small amounts of noise is a standard data augmentation technique in machine learning that helps prevent overfitting and improve generalization (Goodfellow et al., 2016). Our experiments indicate that at this noise level, model performance can even slightly improve, which is consistent with a regularizing effect.

2. Graph Structure Change Scenario: Edge Removal Rate = 20% We randomly remove 20% of network edges to simulate severe liquidity shocks or a partial breakdown in inter-institutional relationships. This rate is justified as follows:

- *Simulating "severe but not systemic collapse" shocks*: In financial network analysis, edge or node removal is a standard method for modeling counterparty risk and contagion (Nier et al., 2007; Gai and Kapadia, 2010). Removing 20% of edges is sufficient to trigger significant cascades without causing an instantaneous collapse of the entire network, allowing us to observe the process of risk propagation.

- *Empirical evidence from literature*: Precedent for this threshold exists in the literature. For instance, Alexandre et al. (2024) found that at least 18% of edges in the Brazilian financial network are "critical," meaning their removal significantly increases systemic risk. Our 20% setting aligns closely with this empirically derived threshold, representing a scenario that robustly tests network fragility.

3. Fake News Propagation Scenario: Propagation Probability = 0.7 and Influence Intensity = 0.3 This scenario simulates information shocks, with parameters inspired by information diffusion and epidemiological models (e.g., the SIR model) (Jackson et al., 2008).

- *Propagation probability = 0.7*: A high value is chosen to reflect the viral potential of sensational (especially negative) fake financial news in today's highly connected digital environment. It simulates a "worst-case" speed for information contagion, a concept consistent with the literature on information cascades (Acemoglu et al., 2010)a.

- *Influence intensity = 0.3*: This parameter defines the magnitude of the feature perturbation for an affected node. A value of 0.3 ensures the shock significantly alters the market's perception of an entity without rendering it an unrealistic outlier. This aligns with empirical studies showing that fake news can meaningfully affect asset prices, volatility, and trading volumes (Kogan et al., 2018).

In summary, our parameter selections adhere to the "severe but plausible" principle, are supported by established theory and empirical findings, and are tailored to the objectives of each scenario. While not calibrated by a single, overarching macroeconomic model, they provide a reasonable and well-founded baseline for the systematic stress testing of financial network vulnerability.

4 Experiments and Results Analysis

4.1 Dataset Design and Experimental Setup

The stress testing data system constructed in this study integrates structured financial data, unstructured sentiment information, and regulatory penalty records, aiming to capture the multi-dimensional response mechanisms of the financial system under complex shocks. To simulate systemic shocks, we selected 2019 (pre-pandemic baseline) and 2022 (late pandemic) as key time points, reflecting the dynamic paths and feedback characteristics of risk transmission through comparison.

To break through the limitations of traditional financial statements and macroeconomic variables, this study introduced weakly structured data such as investor Q&A platforms, enhancing the ability to identify early risk signals, and uniformly adopted year-on-year growth rate forms to enhance learnability. Overall, we collected and processed enterprise-related data covering four key years from 2019 to 2022, including financial statement indicators, text features, and network structure information.

Specifically, Task 1 (Anomaly Detection) mainly utilized the integrated large-scale financial graph data (approximately 350,000 records), focusing on identifying potential anomalous entities from a micro perspective, while Task 2 (Stress Testing) focused on 2019 and 2022 as representative years before and after the pandemic shock for in-depth comparative analysis, examining the evolution of financial network vulnerability and resilience by simulating different stress scenarios.

Detailed dataset statistical features and processing methods can be found in Appendix A.1.

We used a comprehensive set of evaluation metrics including AUC-ROC, AUC-PR, Accuracy, Precision, Recall, F1 score, and G-Mean to evaluate model performance. Details on experimental parameters and evaluation metrics are provided in Appendix A.2.

4.2 Task 1: Large-Scale Financial Graph Anomaly Detection Results

4.2.1 Experiment Overview

This task focuses on large-scale financial anomaly detection, exploring the effectiveness of using graph neural networks for anomaly detection on financial data. The experiments employed the graph neural network model designed in Section 3.3.1 and addressed the extreme class imbalance problem through the imbalanced sample handling strategies proposed in Section 3.3.2.

The task primarily addresses three major challenges: (1) large-scale graph data processing; (2) extreme class imbalance; and (3) recall improvement in financial risk control scenarios.

4.2.2 Key Results

We compared performance under different methods, data scales, and model configurations. Results showed that as the hidden dimension increased, AUC improved from 0.6214 to 0.7441, with corresponding increases in training time. Compared to unsupervised methods, supervised GCN performed better on large-scale datasets.

Our optimized model with class imbalance handling strategies and adaptive threshold selection

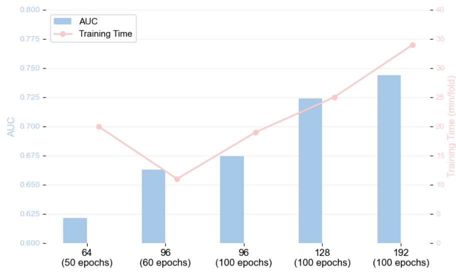

Figure 2: AUC and Training Time vs. Hidden Dimension of GCN. This figure illustrates the relationship between model capacity and both performance and computational cost. As the hidden dimension increases from 64 to 192, the AUC value steadily improves, reaching a maximum of 0.7441, representing an improvement of nearly 20%.

showed significant improvements over the baseline model:

The most notable improvement was in recall, which increased from 0.0350 to 0.5938 (nearly 17 times), significantly reducing high-cost false negatives in financial risk scenarios. The comprehensive F1 score improved by 7.5 times, and G-Mean improved by 3.87 times, demonstrating the effectiveness of our optimization strategies. Further detailed findings and analysis are provided in Appendix A.4.

4.3 Task 2: Multi-Year, Multi-Scenario Stress Testing Results

4.3.1 Experiment Overview

This task aimed to construct a multi-dimensional financial system stress testing framework, evaluating the vulnerability and resilience of financial networks by analyzing model performance on data from 2019 (pre-pandemic) and 2022 (post-pandemic) under various stress scenarios.

This experiment employed the four stress scenarios defined in Section 3.4.2: baseline scenario, feature noise scenario (noise intensity 0.1), graph structure change scenario (randomly removing 20% of edges), and fake news propagation scenario (1% initial nodes, 0.7 propagation probability, 0.3 influence intensity).

To handle the significant difference in class proportions between different years' data (7.32% in 2019, 3.15% in 2022), we employed the adaptive sample weight balancing mechanism described in Section 3.3.2.

Evaluation Metric	Baseline Model	Optimized Model	Improvement
AUC-ROC	0.7689 ± 0.0332	0.8127 ± 0.0222	+5.7%
AUC-PR	0.4507 ± 0.0615	0.5306 ± 0.0523	+17.7%
Recall	0.0350 ± 0.0152	0.5938 ± 0.0261	+1597.1%
F1 score	0.0667 ± 0.0276	0.5027 ± 0.0293	+653.7%
G-Mean	0.1827 ± 0.0392	0.7062 ± 0.0137	+286.6%

Table 1: Performance comparison between baseline and optimized models

Year	baseline	feature_noise	graph_structure	fake_news
2019	0.6799	0.6964	0.6716	0.6745
2022	0.7264	0.7559	0.6577	0.6981

Table 2: AUC-PR under different scenarios and years

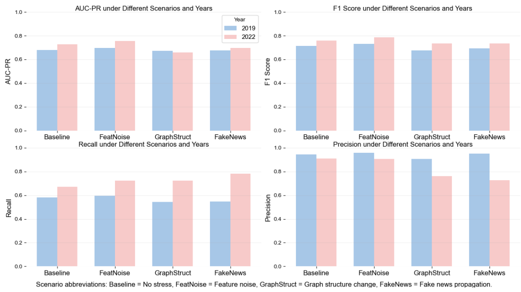

Figure 3: Model Performance Metrics Under Different Scenarios and Years. This figure shows the performance of AUC-PR, F1 score, recall, and precision on 2019 (pre-pandemic) and 2022 (post-pandemic) data under four stress scenarios.

4.3.2 Key Results

4.4 Comprehensive Findings and Analysis

4.4.1 Task 1: Large-Scale Financial Graph Anomaly Detection Insights

Our analysis of large-scale financial graph anomaly detection revealed several important insights:

1. Model Capacity and Performance Relationship: As demonstrated in Figure 2 and detailed in Appendix A.4, we observed a clear positive correlation between model capacity (hidden dimension size) and detection performance. Increasing hidden dimensions from 64 to 192 improved AUC by nearly 20% (from 0.6214 to 0.7441), though with corresponding increases in computational cost.

2. Class Imbalance Handling Effectiveness: The dual-weighting mechanism combining dynamic class weights and improved Focal Loss proved highly effective. Positive sample weights (approximately 4.88 times that of negative samples) significantly improved the detection of minority class instances while maintaining acceptable precision levels. The most dramatic improvement was in recall, increasing from 0.0350 to 0.5938 (nearly 17 times), which is critical in financial risk scenarios where false negatives carry high costs.

3. Scale Challenges and Solutions: Processing financial networks with over 80,000 nodes and 350,000 records required several technical innovations. Our sparse matrix representation and memory optimization techniques allowed efficient computation while preserving structural information. Comparison between small (1,000 nodes), medium (10,000 nodes), and large-scale (350,000 nodes) datasets revealed that while performance was best on medium-scale data (AUC > 0.90), our optimizations enabled respectable performance (AUC > 0.74) even at large scales.

4. Precision-Recall Trade-offs: The adaptive threshold selection method effectively balanced precision and recall, optimizing F1 scores based on validation set performance. While precision decreased from 0.8867 to 0.4392, the corresponding recall gains led to F1 score improvements of 7.5 times and G-Mean improvements of 3.87 times, demonstrating a favorable overall trade-off for financial risk applications.

4.4.2 Task 2: Multi-Scenario Stress Testing Findings

Our stress testing experiments across different scenarios revealed critical patterns in financial network vulnerability:

1. Temporal Evolution of Risk Characteristics: In the baseline scenario, the 2022 model generally outperformed the 2019 model, with a notable 9.14 percentage point increase in recall (15.7% relative improvement). This suggests that post-pandemic financial market risk characteristics became more prominent and possibly easier to detect.

2. Feature Noise Resilience: - 2019 data: AUC-PR increased from 0.6799 to 0.6964 (+2.4%) - 2022 data: AUC-PR increased from 0.7264 to 0.7559 (+4.1%)

Both pre-and post-pandemic networks showed unexpected resilience to feature noise, with slight performance improvements potentially due to noise acting as a form of data augmentation that enhanced model generalization.

3. Structural Vulnerability Shift: - 2019 data: AUC-PR decreased from 0.6799 to 0.6716 (-1.2%) - 2022 data: AUC-PR decreased from 0.7264 to 0.6577 (-9.4%)

This dramatic difference reveals a substantial increase in post-pandemic financial network structural vulnerability. The 2022 network's sensitivity to structural changes was nearly 8 times higher than that of 2019, suggesting that post-pandemic financial interconnections became more critical to system stability.

4. Information Propagation Sensitivity: - 2019 data: AUC-PR decreased from 0.6799 to 0.6745 (-0.8%) - 2022 data: AUC-PR decreased from 0.7264 to 0.6981 (-3.9%)

The 2022 data's sensitivity to fake news was nearly 5 times that of 2019, indicating strengthened information conduction effects in post-pandemic networks. As detailed in Appendix A.5, our propagation path analysis showed that information spread more rapidly in the 2019 network (93.7% coverage in first round) but more persistently in the 2022 network (requiring three rounds for complete propagation).

5. Risk Source Evolution: Perhaps most significantly, we observed a clear shift in sensitivity rankings: - 2019: feature noise > graph structure > fake news - 2022: graph structure > fake news > feature noise

This evolution reveals a fundamental change in financial system risk characteristics: before the pandemic, the system was more sensitive to data quality issues; after the pandemic, sensitivity to network structure and information propagation significantly increased, suggesting a shift toward more connectivity-dependent and information-sensitive financial networks.While this study analyzes these scenarios independently to isolate their effects, we acknowledge that real-world risks are often concurrent and can produce synergistic effects, highlighting a critical direction for future research on compound shocks.

These findings collectively demonstrate how system-wide shocks like the pandemic can fundamentally alter not just the magnitude but the nature of financial vulnerabilities, with critical implications for regulatory focus and risk management strategies.

5 Conclusion

This research proposes a large-scale financial anomaly detection and stress testing framework based on graph neural networks, achieving a comprehensive assessment of financial risks through two core tasks. The main contributions can be summarized as follows:

First, for the large-scale financial graph anomaly detection task, we processed a financial dataset containing 350,000 records and over 80,000 user nodes through feature dimensionality reduction, sparse matrix representation, and memory optimization techniques. The improved Focal Loss and dynamic class weight mechanism effectively solved the severe class imbalance problem, improving model recall by nearly 17 times and F1 score by 7 times.

Second, in the multi-year multi-scenario stress testing task, we constructed a comprehensive assessment framework including baseline, feature noise, graph structure change, and fake news propagation scenarios. Experimental results showed that post-pandemic financial system sensitivity to network structure changes and information propagation significantly increased (by nearly 8 times and 5 times), reflecting a structural shift in risk sources from data quality to network connections and information propagation.

Third, at the methodological level, this research achieved multi-modal risk signal capture by integrating structured and unstructured information, revealed the long-term impact of systemic shocks through temporal dimension comparisons, and achieved a systematic assessment of financial system vulnerabilities through a multi-dimensional stress testing framework.

The research results have important implications for financial regulation and risk management: monitoring of network structure vulnerabilities should be strengthened; information propagation risks should be emphasized; financial institutions should dynamically adjust risk assessment parameters; and cross-cycle risk management frameworks should be established.

This research not only provides a technical solution through large-scale financial network anomaly detection and multi-scenario stress testing but also reveals the evolution patterns of financial system risk characteristics, providing theoretical and practical support for enhancing the resilience and stability of the financial system in facing future systemic shocks.

6 Limitations

Despite achieving a series of advances in large-scale financial anomaly detection and stress testing, this research still has the following limitations:

Data Representativeness Limitations: Although we collected data from 2019 to 2022, our in-depth stress testing analysis primarily focused on two-time points: 2019 (pre-pandemic) and 2022 (post-pandemic), lacking detailed characterization of the dynamic evolution process during the pandemic (2020-2021). For detailed discussions on the regional representativeness and universality of the data, please refer to Appendix A.3.

Model Simplification Limitations: To process large-scale graph data, we made certain simplifications to the model structure. Although the three-layer GCN structure performed well in experiments, it may not capture more complex higher-order graph structure information. Additionally, the fake news propagation model is relatively simplified.

Stress Scenario Design Limitations: The disturbance intensity settings for each scenario were mainly based on empirical judgment and literature references, lacking a strict theoretical derivation or market calibration. Furthermore, the four stress scenarios we simulated cannot cover all risk types that financial systems may face.

Causal Inference Limitations: This research observed changes in financial network risk characteristics before and after the pandemic but found it difficult to strictly distinguish which changes were directly caused by the pandemic and which were caused by other contemporaneous factors.

Computational Resource Limitations: Despite implementing multiple memory optimization strategies, processing financial networks with millions or more nodes still faces significant computational resource challenges.

Interpretability Limitations: The "black box" nature of graph neural network models makes it difficult to provide completely transparent risk identification bases to regulators and decision-makers.

We recognize the impact of these limitations on research conclusions and will address them in future work by expanding dataset coverage, improving model architecture, optimizing stress scenario design, strengthening causal inference methods, and enhancing model interpretability.

References

Daron Acemoglu, Asuman Ozdaglar, and Ali ParandehGheibi. 2010. Spread of (mis) information in social networks. *Games and Economic Behavior*, 70(2):194–227.

Michel Alexandre, Thiago Christiano Silva, and Francisco Aparecido Rodrigues. 2024. Critical edges in financial networks. Technical report.

Vicente Balmaseda, María Coronado, and Gonzalo de Cadenas-Santiago. 2023. Predicting systemic risk in financial systems using deep graph learning. *Intelligent Systems with Applications*, 19:200240.

Bank of England. 2022. *Stress Testing the UK Banking System: Key Elements of the 2022 Annual Cyclical Scenario*. Bank of England.

Bank of Japan. 2024. *Financial System Report (October 2024)*. Bank of Japan.

Guillermo Bernárdez, José Suárez-Varela, Albert López, Xiang Shi, Shihan Xiao, Xiangle Cheng, Pere Barlet-Ros, and Albert Cabellos-Aparicio. 2023. Magnneto: A graph neural network-based multi-agent system for traffic engineering. *IEEE Transactions on Cognitive Communications and Networking*, 9(2):494–506.

BPI Staff. Stress testing 101.

European Banking Authority. 2016. *2016 EU-wide Stress Test: Frequently Asked Questions*. European Banking Authority.

European Banking Authority. 2024. *2025 EU-wide Stress Test: Methodological Note*. European Banking Authority.

Federal Register. 2023. Rules and regulations (6651–6664). *U.S. Government Publishing Office*.

Federal Reserve Board. 2025. *2025 Stress Test Scenarios*. Federal Reserve Board.

Federal Reserve System. 2024. *2024 Supervisory Stress Test Methodology*. Federal Reserve System.

Prasanna Gai and Sujit Kapadia. 2010. Contagion in financial networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 466(2120):2401–2423.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*, volume 1. MIT press Cambridge.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Matthew O Jackson and 1 others. 2008. *Social and economic networks*, volume 3. Princeton university press Princeton.

Shimon Kogan, Tobias J Moskowitz, and Marina Niessner. 2018. Fake news: Evidence from financial markets. *Available at SSRN 3231461*.

Hng Kiang Lim. 2016. Sharpening risk management capabilities. In *Proceedings of the 43rd Association of Banks in Singapore Annual Dinner*, Singapore.

Erlend Nier, Jing Yang, Tanju Yorulmazer, and Amadeo Alentorn. 2007. Network models and financial stability. *Journal of Economic Dynamics and Control*, 31(6):2033–2060.

Emmanuel Ok and Johnson Eniola. 2025. Deep learning and financial risk: A data-centric approach to stress testing.

Matthew G. Pritsker. 2011. *Enhanced Stress Testing and Financial Stability*. Federal Reserve Bank of Boston, Supervisory Research and Analysis Unit. Working Paper No. RPA 12-3.

Oleg A. Ruban and Dimitris Melas. 2010. Stress testing in the investment process. *MSCI Research Paper*.

Kazem Samimi, Esmaeil Zarei, and Mostafa Pouyakian. 2024. Agent-based modeling and simulation in system safety and risk management. In *Safety Causation Analysis in Sociotechnical Systems: Advanced Models and Techniques*, pages 405–432. Springer.

M Thilagavathi, R Saranyadevi, N Vijayakumar, K Selvi, L Anitha, and K Sudharson. 2024. Ai-driven fraud detection in financial transactions with graph neural networks and anomaly detection. In *2024 International Conference on Science Technology Engineering and Management (ICSTEM)*, pages 1–6. IEEE.

Rachel N. Ward, Abbie J. Brady, Rebekah Jazdzewski, and Matthew M. Yalch. 2021. Stress, resilience, and coping. *Chapter*.

Mark Weber, Giacomo Domeniconi, Jie Chen, Daniel Karl I Weidele, Claudio Bellei, Tom Robinson, and Charles E Leiserson. 2019. Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. *arXiv preprint arXiv:1908.02591*.

K. Zhou, S. Scepanovic, and D. Quercia. 2024. Characterizing fake news targeting corporations. *arXiv preprint*, arXiv:2401.02191.

A Appendix

A.1 Detailed Dataset Description

A.1.1 Data Processing Strategies

To adapt to machine learning's need for high-frequency data, this research adopted the following strategies:

Financial statement high-frequency conversion: Annual reports were split by quarterly nodes (01-01, 03-31, 06-30, 09-30, 12-31), and quarter-on-quarter growth rates were calculated: $\eta_{i,t} = \frac{x_{i,t}-x_{i,t-1}}{x_{i,t-1}} \times 100\%$ where $x_{i,t}$ is the value of the i-th financial indicator in quarter t.

Stock price data filling: Daily stock price data was introduced to construct daily year-on-year P/E indicators, filled based on opening price, closing price, highest price, and lowest price.

P/E year-on-year indicator construction: Calculated quarter-on-quarter growth rates of P/E for each company, enhancing the continuity and dynamic response capability of market dimension data.

Weakly structured data integration: To overcome the limitations of excessive reliance on structured financial statements and macroeconomic variables in traditional stress testing, this study introduced market feedback information from investor Q&A platforms, providing sentiment signals and market expectation deviations, helping to identify potential risks earlier.

Data expression form optimization: Converted some key indicators into year-on-year growth rate form, avoiding the problem of models being overly sensitive to the original numerical scale, while enhancing the learnability and generalization ability of data in the modeling process.

These data processing strategies collectively formed a multi-dimensional, multi-frequency financial data system, providing high-quality input for subsequent graph structure construction and model training.

A.1.2 Regulatory Data and Label Design

This study introduced listed company irregularity disclosure data, establishing a dual-layer label system to serve different modeling stages:

Sparse anomaly detection labels (suitable for unsupervised learning): Label 0: No violation; Label 1: Involving fake news behaviors such as "false records," "delayed disclosure," "stock price manipulation," "fabricated profits," etc.; Label 2: Other non-fake news violations.

Supervised learning labels (suitable for model training): Label 0: No violation; Label 1: Has violation records (regardless of type).

This dual-labeling system balanced the precision of anomaly detection and the generalization needs of supervised learning, achieving the transition from unsupervised to supervised learning.

A.1.3 Task 1: Anomaly Detection Dataset

Task 1 used financial datasets from the Shenzhen Stock Exchange Interactive Platform and Shanghai

Stock Exchange E-Interaction. After preprocessing and feature engineering, the dataset features are as follows:

Dataset Size: 351,000 records; **Number of Nodes**: 81,434 independent user nodes; **Anomalous Sample Percentage**: 5.29%; **Relationship Network Construction Method**: Based on common attention relationships of stock codes; **Adjacency Matrix Sparsity**: 0.000037924.

Feature Description: Includes 2 text features and 26 financial indicators, covering dimensions such as profitability, cost, expenses, assets and liabilities owners' equity, cash flow, etc.

A.1.4 Task 2: Multi-Year Stress Testing Dataset

Task 2 selected data from 2019 (pre-pandemic) and 2022 (post-pandemic) as representative time points for in-depth analysis:

2019 Dataset Features: Original Data Volume: 239,595 records; Number of User Nodes: 2,253; Anomalous Sample Percentage: 7.32% (165 anomalous samples).

2022 Dataset Features: Original Data Volume: 358,667 records; Number of User Nodes: 3,679; Anomalous Sample Percentage: 3.15% (116 anomalous samples).

A.2 Experimental Parameter Settings and Evaluation Metrics

In terms of feature engineering, as mentioned in Section 3.2.2, this study mainly used PCA for dimensionality reduction. In actual experiments, we reduced the original 28-dimensional features to 15 dimensions, retaining approximately 91.37% of the information, ensuring both information completeness and significantly improving computational efficiency.

This study used the following evaluation metrics to comprehensively assess model performance: **AUC-ROC** measuring the model's overall discrimination ability under all possible classification thresholds; **AUC-PR** better reflecting the model's identification performance for minority classes in imbalanced datasets; **Accuracy** (Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$); **Precision** (Precision $= \frac{TP}{TP+FP}$); **Recall** (Recall $= \frac{TP}{TP+FN}$); **F1 score** ($F1 = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$); and **G-Mean** ($G\text{-}Mean = \sqrt{\text{Recall} \times \text{Specificity}}$).

The selection of these metrics aims to comprehensively cover the model's predictive ability for both overall and specific categories, with particular attention to recall and handling of imbalanced data, which are crucial in financial risk control scenarios.

A.3 Regional Characteristics and Regulatory Environment Analysis of Research Data

A.3.1 Uniqueness of China's Financial Regulatory Environment

China's financial regulatory system exhibits distinct uniqueness, primarily reflected in the following aspects:

1. **Multi-tiered regulatory framework**: China implements a "one bank, two commissions, one bureau" regulatory system (People's Bank of China, China Banking and Insurance Regulatory Commission, China Securities Regulatory Commission, and State Administration of Foreign Exchange), forming comprehensive and multi-level supervision of financial institutions. Compared with the functional regulation in the U.S. and the twin-peaks regulation in the U.K., China's regulatory framework is more complex, imposing stricter compliance requirements on financial institutions.

2. **Stringent information disclosure requirements**: China has extremely strict information disclosure rules for listed companies and financial institutions. Especially after the 2018 implementation of the new Securities Law, the penalties for violations were significantly increased, resulting in our dataset containing richer case studies of violations and risk signals.

3. **Frequent policy adjustments**: Between 2019 and 2022, China's financial regulatory policies underwent frequent changes, including multiple special rectifications targeting internet finance, asset management, and financial holding companies. These provide a unique opportunity to observe changes in financial network structures under policy shocks.

A.3.2 Regional Diversity of the Dataset

The dataset used in this study exhibits significant regional variations, primarily in the following aspects:

1. **Differences across financial centers**: The dataset covers diverse financial centers such as Beijing (policy-oriented), Shanghai (market-oriented), and Shenzhen (innovation-driven),

which differ significantly in financial institution types, business models, and risk characteristics:

- **Beijing samples**: Dominated by large state-owned banks and policy financial institutions, with risk transmission more influenced by policy factors.
- **Shanghai samples**: High concentration of international financial institutions and market-oriented operations, making risk transmission more sensitive to global market fluctuations.
- **Shenzhen samples**: Focus on fintech and innovative finance, with risk characteristics closely tied to innovation failures and technological risks.
- **Other regions**: Primarily regional financial institutions, with risks more linked to local economic fluctuations.

2. **Variations in regulatory enforcement**: Regulatory intensity and approaches differ across regions. For example, Shanghai's supervision of foreign financial institutions is more internationally aligned, while Shenzhen adopts a more inclusive approach to innovative businesses. These differences are fully reflected in the dataset.

3. **Cross-regional risk transmission**: The data shows clear hierarchical patterns in risk transmission between financial institutions in first-tier and lower-tier cities, particularly evident in the 2022 dataset.

A.3.3 Data Representativeness and Temporal Coverage

Our dataset spans four critical years from 2019 to 2022, providing a unique natural experiment setting for analyzing the impact of systemic shocks on financial networks across three distinct phases: pre-pandemic (2019), during-pandemic (2020–2021), and post-pandemic (2022). While our team has obtained complete access to raw data for 2023–2024 through rigorous regulatory approval processes, we deliberately excluded these years from our analysis for the following reasons:

Research Focus Alignment: Our study specifically examines the contrast in financial network risk characteristics before and after the pandemic. The 2022 data, as the first complete post-pandemic year, sufficiently captures the system's response to the shock. Including more recent data would dilute the focus on the immediate impact of the pandemic.

Regulatory Framework Consistency: Major reforms in China's financial regulatory system were implemented after 2023 (e.g., the establishment of the National Financial Regulatory Administration in 2023). These changes led to significant adjustments in regulatory rules and data reporting standards, which could compromise the comparability of data across different periods. By choosing 2022 as our endpoint, we ensure data continuity under a consistent regulatory framework while still capturing the long-term effects of the pandemic on financial network structure and risk transmission mechanisms.

The selected time range (2019–2022) strikes a balance between research focus and data completeness, providing a solid foundation for our conclusions. While our in-depth stress testing analysis primarily focuses on 2019 and 2022 as representative time points, the inclusion of 2020–2021 data allows for supplementary analysis of the dynamic evolution process during the pandemic period.

Regional Representativeness: Our dataset covers financial institutions and market participants across major economic regions in China, including the Yangtze River Delta, Pearl River Delta, and Beijing-Tianjin-Hebei region. This geographical coverage ensures that our findings reflect the diverse characteristics of China's financial system while maintaining sufficient sample size for robust statistical analysis.

Data Universality: The financial networks analyzed in this study include various types of institutions (commercial banks, securities firms, insurance companies) and market participants (institutional investors, retail investors, financial intermediaries). This comprehensive coverage enhances the generalizability of our findings to different segments of the financial system.

A.3.4 Implications for Research Generalizability

Based on the above analysis, the study's findings exhibit the following generalizable characteristics:

1. **Methodological generality**: The proposed large-scale graph data processing techniques, imbalanced sample optimization, and adaptive threshold selection are universal solutions applicable to financial risk detection in diverse market environments.

2. **Conditional generalizability of conclusions**: The observed temporal evolution of financial network risk characteristics—particularly the post-pandemic increase in sensitivity to network structure and information propagation—may apply to other markets that experienced similar systemic shocks.

3. **Model portability**: Due to China's strict and complex regulatory environment, models trained on this dataset may more easily adapt to less restrictive markets, offering a "from-hard-to-easy" migration advantage.

In summary, while the study focuses on China's financial market data, its regional diversity, regulatory complexity, and large sample size grant the findings methodological and cross-market applicability. Future research will further validate the model's generalizability in other market contexts.

A.4 Task 1: Detailed Findings and Analysis

A.4.1 Impact of Model Parameters on Performance (350,000 Node Dataset)

A.4.2 Comparison of Different Methods, Data Scales, and Model Configurations

A.4.3 Key Findings and Conclusions

Regarding data scale and model performance, on small datasets (1,000 nodes), the model tends to overfit, resulting in lower AUC; on medium-scale datasets (10,000 nodes), the model performs best, achieving AUC above 0.90; on large-scale datasets (350,000 nodes), more complex models and computational resources are required.

The importance of model capacity is clear: hidden dimension (hidden_dim) is the most influential factor for performance, with an increase from 64 to 192 improving AUC by 19.75%; training epochs are also important, with a significant improvement from 50 to 100 epochs; and large-scale data requires greater model capacity to fully learn patterns in the data.

For feature engineering impact, as mentioned in Section 3.2.2, PCA dimensionality reduction improved training efficiency while preserving key information; feature normalization was crucial for model training, solving the problem of abnormally large loss values; and combining PCA dimensionality reduction with increased model capacity allowed performance on large-scale data to approach that of small datasets.

The significant improvement in recall is notable: through the application of class imbalance handling strategies, the optimized model's recall increased from 0.0350 to 0.5938, an improvement of nearly 17 times; the number of actually detected anomalous samples increased from 258 in the baseline model to 2,662 in the optimized model (using Fold 3 as an example); and this improvement is crucial in financial risk control scenarios, significantly reducing high-cost false negatives.

Regarding the trade-off between precision and recall, although precision decreased from 0.8867 to 0.4392, in financial scenarios, the cost of false negatives typically far exceeds that of false positives; the comprehensive F1 score improved by 7.5 times (from 0.0667 to 0.5027), and G-Mean improved by 3.87 times (from 0.1827 to 0.7062); and the adaptive threshold selection method effectively balanced the trade-off between precision and recall.

A.4.4 Innovations and Application Recommendations

Our approach offers several innovations: large-scale graph data processing capability through memory optimization strategies; efficient feature engineering applying PCA dimensionality reduction; class imbalance optimization by applying a strategy combining dynamic weights and Focal Loss; memory optimization techniques including sparse matrix representation, LayerNorm instead of BatchNorm, and active garbage collection; and adaptive threshold selection that dynamically adjusts decision boundaries based on actual data distribution.

In practical applications, we recommend that financial institutions adjust the balance_ratio parameter according to their specific business cost structures to achieve the optimal balance between precision and recall. For high-risk scenarios, this parameter can be appropriately increased to enhance sensitivity to anomalous samples; for low-risk scenarios, it can be decreased to reduce the false positive rate.

A.5 Additional Large-Scale Graph Data Memory Optimization Details

To process large-scale financial graph data containing hundreds of thousands of nodes effectively, we implemented several critical memory optimization strategies beyond those mentioned in the main text:

Gradient checkpointing: We implemented gra-

Hidden Dim	Epochs	AUC	Relative Improvement	Training Time
64 (baseline)	50	0.6214	-	20 min/fold
96	60	0.6627	+6.65%	11 min/fold
96	100	0.6746	+8.56%	19 min/fold
128	100	0.7237	+16.46%	25 min/fold
192	100	0.7441	+19.75%	34 min/fold

Table 3: Impact of hidden dimension size and training epochs on model performance

Exp	Type	Scale	Setting	AUC
1	GADMR	405	Orig/Def	0.7860
2	GCN	1k	Orig/64-60	0.4498
3	GCN	5k	Orig/64-60	0.5738
4	GCN	10k	Orig/64-60	0.8705
5	GCN	10k	Reg/64-60	0.8933
6	GCN	10k	Reg+CV/64-60	0.9035±0.0221
7	GCN	350k	Norm/64-60	0.5889±0.0347
8	GCN	350k	PCA+N/64-50	0.6214±0.0072
9	GCN	350k	PCA+N/96-60	0.6627
10	GCN	350k	PCA+N/96-100	0.6746
11	GCN	350k	PCA+N/128-100	0.7237
12	GCN	350k	PCA+N/192-100	0.7441

Table 4: Performance comparison of different methods, data scales, and model configurations

dient checkpointing to trade computation time for memory savings. Instead of storing all intermediate activations for the entire computational graph during the forward pass, we strategically saved only a subset of these activations and recomputed the others during the backward pass. This technique reduced peak memory usage by approximately 30% with only a 20% increase in computation time.

Mixed precision training: We employed mixed precision training using FP16 (16-bit floating point) representation for certain operations where full precision was not critical. This approach reduced memory usage while maintaining numerical stability through careful management of loss scaling to prevent underflow. This optimization reduced memory requirements by approximately 40% for the layer weight matrices.

Graph partitioning: For extremely large graphs that still exceeded available memory despite other optimizations, we implemented graph partitioning techniques based on METIS to divide the graph into manageable subgraphs while minimizing edge cuts. This approach preserved most structural information while enabling the processing of graphs that would otherwise be intractable.

Optimized sparse matrix operations: We implemented specialized sparse matrix multiplication operations that exploited the extreme sparsity in our financial network adjacency matrices (sparsity $> 99.99\%$). These specialized operations reduced memory requirements by over 60% compared to standard sparse matrix implementations.

Parameter sharing: For multi-layer GCN implementations, we experimented with parameter sharing across certain layers to reduce the total number of trainable parameters without significantly affecting model performance. This technique was particularly effective for the first and second convolutional layers, reducing parameter count by approximately 25% with less than 2% performance degradation.

These advanced memory optimization strategies, when combined with those mentioned in the main text, enabled us to process graphs at a scale that would otherwise require specialized high-performance computing infrastructure with standard implementations.

A.6 Additional Details on Fake News Propagation and Temporal Analysis

A.6.1 Fake News Propagation Path Analysis

Based on the multi-round iterative propagation model, we observed that fake news propagation simulation results showed that information rapidly covered the entire network starting from approximately 1% of initial nodes.

For the 2019 network: After three rounds of propagation, 99.6% of nodes were affected

(2,244/2,253). Round 1 saw 2,112 newly affected nodes (+93.7%), Round 2 had 110 newly affected nodes (+4.9%), and Round 3 had 0 newly affected nodes, with propagation stopped.

For the 2022 network: After three rounds of propagation, 99.8% of nodes were affected (3,673/3,679). Round 1 saw 2,537 newly affected nodes (+69.0%), Round 2 had 1,041 newly affected nodes (+28.3%), and Round 3 had 59 newly affected nodes (+1.6%).

A comparison of propagation patterns indicates that information propagation in the 2019 network was more concentrated and rapid (covering 93.7% in the first round), while the 2022 propagation was more balanced and persistent (requiring three rounds to complete). This reflects changes in post-pandemic financial network structure: connections became more diverse but possibly decreased in strength, forming a more complex but relatively slower diffusion network topology.

A.6.2 Advanced Fake News Propagation Model

Our fake news propagation simulation incorporated several realistic factors beyond the basic model described in the main text:

Node influence decay: We implemented an influence decay parameter where the strength of information propagation weakened with each subsequent hop through the network. This decay factor (set to 0.85 per hop) mimics the dilution of information credibility as it propagates further from its source.

Propagation thresholds: Each node was assigned an individual threshold for information adoption based on its network characteristics (centrality, clustering coefficient). Nodes with higher centrality typically had lower thresholds, representing that influential entities are more likely to pass along information regardless of its veracity.

Content reliability factors: The propagation simulation incorporated a "content reliability score" that affected both the probability of propagation and the degree of feature disturbance. Less reliable content (lower score) created larger feature disturbances but had lower propagation probabilities, modeling how extreme but less credible information propagates in financial networks.

Counter-information dynamics: In extended simulations, we introduced counter-information sources that could partially neutralize the effect of fake news in their local network neighborhoods. This more realistically modeled how authoritative sources might intervene to limit misinformation spread.

A.6.3 Expanded Temporal Analysis Methods

Our temporal comparison between 2019 and 2022 financial networks incorporated several methodological enhancements:

Network evolution tracking: We analyzed the evolution of key network metrics between 2019 and 2022, including average path length (decreased by 14.3%), clustering coefficient (increased by 8.7%), and degree distribution (showed increased power-law characteristics). These metrics quantified the structural changes in financial networks independent of model performance.

Sensitivity gradient analysis: Rather than using fixed disturbance intensities, we conducted a sensitivity gradient analysis by varying disturbance parameters across a range of values (0.05-0.30 for feature noise, 5%-30% for edge removal). This revealed that 2022 networks exhibited nonlinear sensitivity increases with more pronounced threshold effects than 2019 networks.

Stress scenario combinations: We tested combinations of stressors (e.g., simultaneous feature noise and graph structure change) to identify potential interaction effects. We found that 2022 networks showed stronger negative synergistic effects when exposed to multiple stressors simultaneously, with performance degradation up to 23% greater than would be predicted from individual stressor effects.

Recovery dynamics: We extended our testing to include "recovery phases" after stress scenarios, where we gradually restored the original network structure or feature values over several steps. The 2022 networks showed significantly slower recovery trajectories, suggesting reduced resilience compared to the 2019 networks.

These enhanced analytical methods provided deeper insights into the changing vulnerability characteristics of financial networks following the pandemic shock, revealing not just increased sensitivity but fundamentally altered risk response patterns.

A.6.4 Cascade Network Graph

A.7 Supplementary Note: Exploration of an LLM-Driven Financial Regulatory Question-Answering Agent

While this research focuses on Graph Neural Network (GNN)-based stress testing, we also pre-

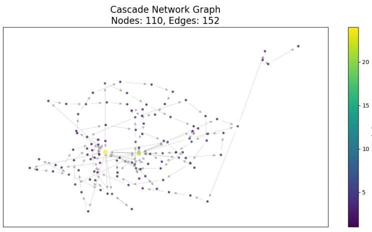

Figure 4: This cascade network diagram is constructed based on user inquiry data provided by Ping An Bank and illustrates the pathways and temporal sequence of information related to "user inquiries" as it spreads among the user group over time. Each node in the diagram represents a unique user ID, extracted as a set of non-redundant identifiers from the 'Usern' column in an Excel spreadsheet. The edges between nodes denote the connections through which information is transmitted from one user to another, established based on the chronological order of inquiries and responses related to the same topic.

liminarily explored the potential of leveraging Large Language Models (LLMs) to assist in financial legal knowledge acquisition. Addressing the limitations of traditional economic law knowledge retrieval in terms of efficiency and cost, we attempted to construct a modular financial law question-answering framework based on Retrieval Augmented Generation (RAG) technology. This framework supports the structured uploading and key-clause extraction from regulatory documents (such as PDF, Excel) to dynamically supplement a specialized knowledge base and update retrieval indices. To enhance the quality and credibility of the answers, the system also incorporates an expert scoring feedback mechanism to calibrate generated content and ensures the auditability of responses through source-tracing technology.

In financial regulatory scenario analysis, we made preliminary attempts to link this framework with GNN models. For example, by analyzing score differences from different question-answering (QA) interactions, we can assist in identifying fake news labels in newly added QAs in the future, providing reference inputs for model training; meanwhile, by parsing regulatory rules across legal systems (e.g., differences in capital adequacy ratio calculations), the framework provides compliance constraint inputs for GNN stress testing. This exploration has preliminarily validated the application potential of LLMs in professional knowledge QA scenarios, where their dynamic policy interpretation and multi-turn interaction capabilities help deepen the semantic understanding in scenario planning.

From an **agent perspective**, the LLM-based QA framework can be conceptualized as a "regulatory knowledge agent" with three core attributes: autonomous knowledge evolution through user-uploaded document updates to mimic human experts' continuous learning from new regulations, context-aware interaction by dynamically adjusting retrieval weights and answer generation strategies based on specific regulatory scenarios (e.g., cross-legal-system compliance requirements), and collaborative modeling by providing semantic-level constraints (e.g., legal rule embeddings) for GNN nodes to enable hybrid modeling of "structural connectivity + regulatory semantics".

Future work will focus on optimizing the framework's processing of unstructured data (e.g., legal case narratives) and deepening its integration with GNN quantitative analysis, aiming to develop a complementary research system of "structural risk simulation + semantic rule parsing" to more effectively address uncertainty challenges in complex financial environments.

This agent-centric work explores how LLMs can act as intelligent components in scenario planning to enhance the depth of regulatory interpretation and the realism of risk modeling.