Hindawi
Journal of Control Science and Engineering
Volume 2022, Article ID 8401354, 11 pages
https://doi.org/10.1155/2022/8401354

Research Article

On the Effectiveness of Graph Statistics of Shareholder Relation Network in Predicting Bond Default Risk

Zhiguo Huang [1,2]

[1] *Sci-Tech Academy, Zhejiang University, Hangzhou 310058, China*
[2] *Post-Doctoral Research Center, Hundsun Inc, Hangzhou 310053, China*

Correspondence should be addressed to Zhiguo Huang; hzg0601@163.com

Received 25 August 2022; Revised 14 October 2022; Accepted 22 November 2022; Published 30 November 2022

Academic Editor: Sundarapandian Vaidyanathan

Starting from the theoretical effectiveness of shareholder relation network information for predicting bond default risk, we propose two efficient schemes for extracting two different graph statistics of shareholder relation networks: graph structure statistics and graph distance statistics. In order to test the effectiveness of the two schemes, seven machine learning methods and three types of prediction tasks are used. The shareholder relation network information's effectiveness and machine learning methods are also analyzed. Results show that the graph statistics of shareholder relationship networks are insufficient to be used independently as input features for predicting bond default risk but can provide helpful incremental information based on financial features. The shareholder relation information is effective for predicting bond default risk. The structure statistics perform best among all graph statistics overall, and Cascade Forest and LightGBM perform best among all seven machine learning methods.

1. Introduction

Since 2016, China's bond default risk has begun to be exposed at an accelerated rate, and bond default events have occurred frequently, adversely affecting the bond market's direct financing function. With the accelerated credit risk exposure, the effective prediction of bond default risk is of great significance to bond issuers, investors, underwriters, rating agencies, and regulators.

However, traditional bond default risk predicting methods have limited predicting power. Therefore, new features and new models are urgently needed. Depending on the data source, traditional methods can be roughly divided into two classes: one based on financial data and another based on historical price or return. Methods based on financial data usually select a certain number of financial indicators based on expert experience and use machine learning methods as prediction models to treat default events as a classification problem. The methods based on historical price or return treat the default risk prediction problem as a time series analysis problem and analyze the default risk by predicting return or price through econometric models. Traditional methods only using financial data and historical prices or returns cannot capture comprehensive information for predicting default risk. A "shareholder relation network" is a kind of information that plays an important role in default events but has not been effectively utilized.

The shareholder relationship network has theoretical significance for predicting bond default risk. A "shareholder relation network" refers to the network of shareholders and shareholding ratios. It is the basis of corporate governance, which determines the organizational structure of an enterprise, thus further determining the behavior pattern of the enterprise. The behavior pattern of the enterprise is often the internal incentive for bond default events. Additionally, when the bond default risk occurs, the shareholder relation network also determines the willingness of shareholders and related parties to rescue, thereby determining whether the default events are triggered or not. Therefore, the shareholder relationship network has theoretical significance for predicting bond default risk. However, there are still challenges in effectively extracting information from shareholder relation network graphs due to the lack of efficient and web-scaled graph statistics solutions.

In the long-term research history of graph information extraction, a series of graph statistics has been defined to measure the structural properties of a node or graph, such as degree, clustering coefficient, and transitivity. However, the time complexities of those graph statistics are different. For example, let $\mathbf{V}$ denote the node set and $\mathbf{E}$ denote the edge set, the time complexity of the betweenness centrality is $O(|\mathbf{V}| * |\mathbf{E}|)$. Given that the number of target nodes is d, the total time complexity is $O(d * |\mathbf{V}| * |\mathbf{E}|)$, which is acceptable for small-scale graphs but disastrous for web-scaled graphs. How to efficiently extract the structural information of a web-scaled graph is the key point for applying graph structure statistics. In addition, with the development of graph neural networks, the consensus has been reached that the distance distributional information of nodes in the local neighborhood subgraph (i.e., ego-net) to the target node is of great significance for measuring the importance of nodes. However, a unified, efficient local neighborhood subgraph distance distributional statistic extraction scheme is yet lacking. For this reason, we propose two efficient schemes for extracting statistics from shareholder relationship networks: the graph structure statistic extracting scheme and the graph distance statistic extracting scheme. The time complexities of both schemes are $O((|\mathbf{V}_{\mathbf{m}}| + |\mathbf{E}|) * d)$, where $\mathbf{V}_{\max}$ denotes the node set of the largest local neighborhood subgraph in the target node.

Graph statistics of shareholder relationship networks provide us with new features for predicting bond default risk, and the rapid development of machine learning provides us with new prediction models. Recently, machine learning algorithms for tabular data have made great progress, but their effectiveness in predicting bond default risk has not been tested. Thereby, we test several state-of-the-art machine learning algorithms for tabular data as the secondary goal of the work.

We claim that if one kind of feature can be used as an independent input feature to predict bond default risk or provide incremental information about predicting bond default risk based on financial features, then the kind of feature is said to be effective. If one kind of feature performs better than other kinds of features, then the kind of feature is said to be relatively effective. In order to test the relative and absolute effectiveness of different types of information and the effectiveness of different machine learning algorithms for tabular data in predicting bond default risk, we conduct extensive experiments. Our contributions are concluded as follows:

(1) We propose an efficient scheme for extracting the graph structure statistics of the shareholder relation network

(2) We propose an efficient scheme for extracting graph distance statistics of the shareholder relation network

(3) Based on the seven prediction models and three types of prediction tasks, we test the relative effectiveness and absolute effectiveness of two kinds of graph statistic extraction schemes

(4) We conduct extensive experiments to test the effectiveness of various machine learning algorithms for tabular data in bond default risk prediction

2. Related Works

The literature related to this work mainly focuses on the following four aspects: machine learning prediction models for tabular data; the effectiveness of shareholder relationship networks in bond default or credit risk prediction; the effectiveness of graph structure information or graph statistics in financial risk management; and the effectiveness of local neighborhood subgraph distance information in capturing node topology information.

2.1. Machine Learning Prediction Models for Tabular Data. The two basic machine learning models for tabular data predicting bond default risk are decision tree [1] and MLP [2]. The decision tree is the basis of ensemble learning, and MLP is the basis of deep learning. Ensemble learning methods can be divided into three classes: bagging, boosting, and stacking. The mainstream algorithm that belongs to bagging is random forest [3]. It is an instance of bagging based on the decision tree, which contains two mechanisms: random sample selection and random attribute selection. With those two mechanisms, a balance of precision and efficiency is achieved. The mainstream algorithms that belong to boosting are GBDT [4], Xgboost [5], and LightGBM [6]. The GBDT is the first boosting algorithm, which improves by serially reducing the classification error and fitting the classification residuals of the previous learners with a new learner iteratively. The Xgboost has made various improvements to solve the problems of slow training speed and accuracy of GBDT and greatly improved the training speed and prediction performance. LightGBM is another improved GBDT based on the GOSS (gradient-based one-side sampling) and the EFB (exclusive feature bundle) to achieve a balance between accuracy and efficiency. The mainstream algorithm for stacking is Deep Forest [7]. It is an algorithm that stacks different types of forests in width and depth, consisting of two modules: the first is used to reflect the difference in input data and the second is used to improve the classification ability in input data. The former is called multigranularity scanning, and the latter is called Cascade Forest. See Dong et al. [8] for more ensemble learning algorithms.

2.2. The Effectiveness of Shareholder Relation Network in Bond Default or Credit Risk Prediction. Gantchev and Chakraborty [9] studied the relationship between shareholder relationship networks and bond issuance in the private placement, which showed that a good equity structure and shareholder relationship network could reduce the bond default risk. King [10] studied the impact of corporate governance structure on managers' investment strategies, which showed that a strong shareholder governance structure would make company management adopt low-risk investment strategies such as capital expenditures, while a weak shareholder governance structure would make management adopt high-risk investment strategies such as R&D investment, which thereby triggers the default risk. Garlappi et al. [11] introduced the role of shareholders in studying the correlation between stock return and bond default probability, which showed that shareholder advantage significantly impacts bond default probability. Shi et al. [12] examined the correlation between the controlling shareholder's equity pledge and the protection of corporate creditors' interests based on the data from China's bond market, which showed that the potential control transfer risk of equity pledges could easily lead to the opportunistic behavior of controlling shareholders, which infringes on the interests of corporate creditors. Wu et al. [13] studied the impact of controlling shareholder's equity pledges on bond credit spread. In summary, information about shareholder relationship networks has theoretical significance for predicting bond default or credit risk. However, the application of it based on graph statistics is still relatively few.

2.3. The Effectiveness of Graph Structure Information or Graph Statistics in Financial Risk Management. Yıldırım et al. [14] extracted features from the intercompany transaction network to predict corporate default and achieved good performance. Lee et al. [15] used the graph convolutional neural network and the virtual distance of the debtor in the network for credit default prediction and achieved good performance. Lu et al. [16] used multilayer and parallel connected graph convolutional neural networks for default prediction in P2P networks. Lv et al. [17] analyzed the risk contagion problem in the guarantee network. Lee et al. [15] used the Internet financial lending network to evaluate the credit of users applying for lending and achieved higher accuracy than the BP neural network. Li and Zhang [18] formed a directed weighted network based on equity and related transactions and combined it with the SIRS model to study the risk propagation mechanism within a group. Qian and Xu [19] used a two-layer network to describe the coupling network formed by entrepreneurs and enterprises and analyzed the influence path of entrepreneurs' social relationships on associated credit risk contagion. Wang and Zhang [20] implemented credit default early warning modeling by constructing a GR-LDA method containing a graph structure. Zhang and Li [21] analyzed the network structure features of enterprise groups based on complex network theory. They conducted empirical analysis on the contagion and risk spillover effects of enterprise group credit risk. In summary, graph information or graph statistics could effectively manage financial risk. However, the use of graph statistics for shareholder relationship networks to predict bond default risk is still relatively few.

2.4. The Effectiveness of Local Neighborhood Subgraph Distance in Capturing Node Topology Information. By adding the spatial information of nodes to the Transformer, Ying et al. [22] enabled the Transformer to surpass existing SOTA GNNs models and won the OGB large-scale challenge 2021 championship. By adding location information, You et al. [23] achieved significantly improved performance over the baseline model. By adding distance information, Li et al. [24] improved the performance of the baseline model. Alsentzer et al. [25] achieved state-of-the-art performance on the subgraph regression task by adding neighborhood, structure, and location information. To sum up, the local neighborhood subgraph distance information has theoretical significance in capturing the local topology information of nodes. However, there are few solutions for capturing graph distance information-based on graph distance statistics.

3. Methodology

3.1. Overall Framework. The framework of the work consists of 5 steps:

(1) Generating financial features and labels based on financial expert experience
(2) Generating graph statistics features of the shareholder relation network
(3) Selecting machine learning algorithms
(4) Training machine learning algorithms
(5) Analyzing the predicting performance of features and machine learning algorithms

The overall architecture of the work can be depicted in Figure 1.

3.1.1. Generating Financial Features and Labels Based on Expert Experience. Although shareholder relation networks can theoretically provide incremental information for bond default risk prediction, the trigger of bond default risk is not only determined by it. Traditionally, financial features play an important role in the prediction of bond default risk. Therefore, we first extract 89 financial indicators based on expert experience as benchmark features. Based on the predicting performance improvement based on those financial features, we can analyze the effectiveness of the graph statistics of the shareholder relations network.

Although bond default risk prediction can traditionally be regarded as a classification task for bond default event prediction, the number of default bonds in China's bond market is still few enough, so the labels for the classification task are relatively sparse. It is difficult to comprehensively measure the effectiveness of the graph statistics of a shareholder relationship graph under the classification task of a bond default event. In practice, the practitioners in the bond market often use abnormal changes in return as an agent indicator of bond default risk. Since bonds are a type of fixed income asset, they tend to have a relatively low return. Therefore, when the underlying return exceeds 8%, the bond is usually considered at risk of default. As a result, the bond return can be used as the label to indicate the default risk, and the bond default risk prediction is transformed into a regression task. In order to comprehensively measure the

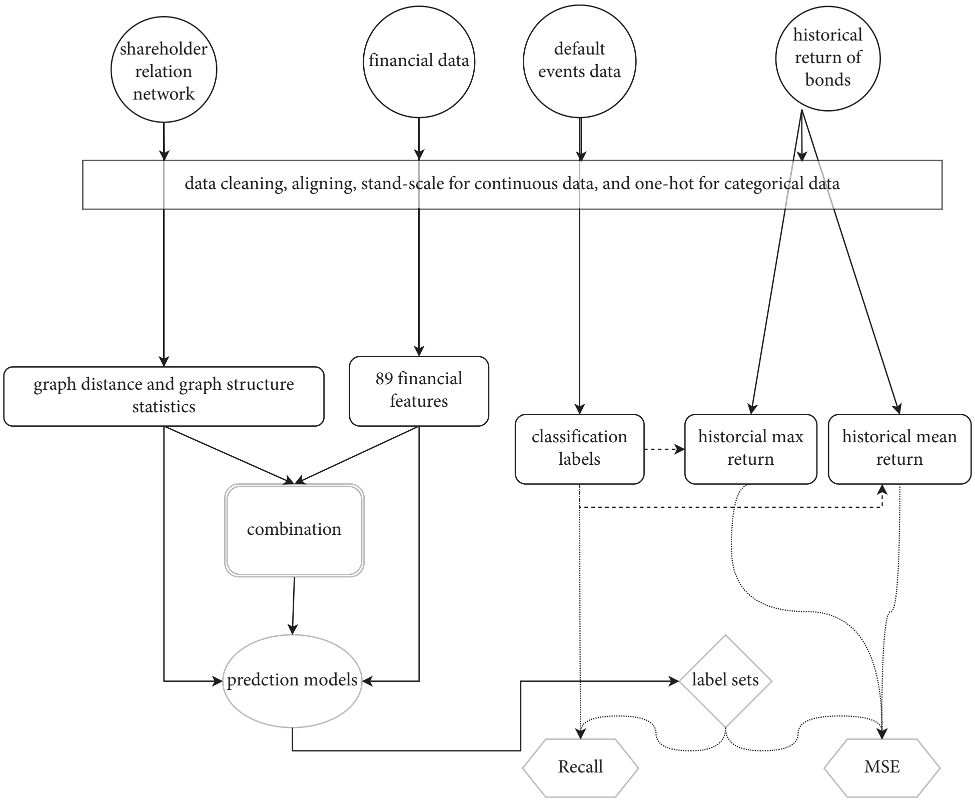
shareholder relation network
financial data
default events data
historical return of bonds
data cleaning, aligning, stand-scale for continuous data, and one-hot for categorical data
graph distance and graph structure statistics
89 financial features
classification labels
historcial max return
historcial mean return
combination
predction models
label sets
Recall
MSE

Figure 1: Overall architecture.

effectiveness of the graph statistics of the shareholder relation network, in addition to classification labels constructed by bond default events, we construct two types of regression task labels: the historical maximum return and the historical average return. In addition, to effectively utilize the information of real bond default events, we set the return of real default bonds at 50% to augment regression labels.

3.1.2. Generating Statistical Features of the Shareholder Relation Network. The graph statistics of shareholder relation networks in this work are divided into graph structure and distance statistics. Although extracting node structure graph statistics based on the entire graph can capture more global information, the time complexity of this pattern is too high to apply to large-scale graphs. In order to reduce the time complexity and extract as multigranular structural information as possible, we propose to use a node-based local neighborhood subgraph to extract graph statistics. The extracting scheme consists of two types of graph structure statistic types: node level and graph-level. Node level statistics include degree, triangles, clustering, and square clustering, and graph-level statistics include density, number of self-loops, and transitivity.

Another type of graph statistic we extract is the distance statistics of local neighborhood subgraphs. The extracting scheme consists of two types of distance statistics: distance distribution statistics and distance count statistics. We selected seven distance distribution statistics: maximum, minimum, median, mean, standard deviation, kurtosis, and skewness. The count statistic is the number of nodes within a given hop around the target node.

3.1.3. Selecting Machine Learning Algorithms. Machine learning algorithms are used to predict bond default risk based on raw financial data, graph distance statistics, and graph structure statistics. Since we choose to use bond default risk as both a classification task and a regression task, the criterion for selecting an algorithm is that it can perform both classification and regression tasks. The type of model is also an important criterion for the consideration of the diversity of models. For those reasons, we select six algorithms in four categories: Random Forest, which belongs to bagging ensemble learning; Cascade Forest, which belongs to stacking ensemble learning; GBDT, Xgboost, and LightGBM, which belong to boosting ensemble learning; and MLP and TabNet, which belong to neural networks. The details of each algorithm can be found in subsection 3.3.

3.1.4. Training Machine Learning Algorithms. Following the usual paradigm of training machine learning algorithms (in this work, we use "machine learning algorithms," "prediction model," and "algorithm" interchangeably), we first divide the data into a train set and a test set, training models on the train set and evaluating the performance of features and models on the test set. The goal of model training is to evaluate different features, feature combinations, and algorithms, so it is necessary to traverse them. The traversal order for the feature and feature combination selection is (1) finance features; (2) graph structure statistics; (3) graph distance statistics; (4) finance + graph distance statistics, (5) finance + graph structure statistics; and (6) finance + graph structure + graph distance feature. The traversal order for the prediction model is consistent with the prediction model selection: (1) Random Forest, (2) Cascade Forest, (3) GBDT, (4) Xgboost, (5) LightGBM, (6) MLP, and (7) TabNet. We decompose the work into 63 training and prediction tasks using the above feature traversal and prediction model strategies.

3.1.5. Analyzing the Test Performance of Features and Machine Learning Algorithms. Based on the test performance, we can systematically analyze the following target: (1) the absolute effectiveness of the different types of graph statistics in predicting bond default risk; (2) the relative effectiveness of the different types of graph statistics in predicting bond default risk; (3) the overall effectiveness of the shareholder relation network in predicting bond default risk; and (4) the effectiveness of each machine learning algorithm in the prediction of bond default risk.

3.2. Extracting Graph Statistics of Shareholder Relation Network. The overall extraction schema can be depicted in Figure 2.

The procedure for extracting graph statistics from the shareholder relation network is listed as follows:

(1) For every node in the graph, extract $k-$hop local neighborhood subgraph around it, where $k = 1, 2, 3, \ldots, K$ and K is a user-specific integer parameter;

(2) For every $k-$hop local neighborhood subgraph, extract its node level and graph-level structure statistics.

  (a) Node level structure statistics consists of degree, triangles, clustering, and square clustering, which are computed as follows:
  Degree: the number of one-hop neighbors of the target node.
  Triangles: the number of triangles around the target node.
  Clustering: $C_u = 2T(u)/\deg(u)(\deg(u) - 1)$, where $T(u)$ denotes the number of nodes passing through a node u and $\deg(u)$ denotes the degree of the node u.
  Square clustering: $C_4(v) = \sum_{u=1}^{k_v} \sum_{w=u+1}^{k_v} q_v(u,w) / \sum_{u=1}^{k_v} \sum_{w=u+1}^{k_v} [a_v(u,w) + q_v(u,w)]$, where $q_v(u,w)$ denotes the number of common neighbors of u, v except v, and $a_v(u,w) = (k_u - (1 + q_v(u,w) + \theta_{uv})) + (k_w - (1 + q_v(u,w) + \theta_{uw}))$, where $\theta_{uw} = 1$ if u, w are connected, otherwise 0.

  (b) Graph-level structure statistics consists of following statistics: density, self-loops and transitivity, which are computed as follows:
  Density: $d = 2m/n(n-1)$, where n denotes the number of nodes and m denotes the number of edges.
  Self-loops: the number of self-loops in the graph.
  Transitivity: $T = 3\#\text{triangles}/\#\text{triads}$, triad denotes a connected graph with two edges.

(3) For every $K-$hop local neighborhood subgraph, extract normalized distance distribution statistics.

  (a) The distance distribution statistics include maximum, minimum, median, mean, standard deviation, kurtosis, skewness, and count.

  (b) Since the local neighborhood of each node is not Euclidean, the number of nodes in each $K-$hop is different, so the dimension of every series of distance distribution statistics may be different. In order to make each series the same dimension, we implement zero pads for every extracted series.

3.3. Selecting Machine Learning Algorithms. As mentioned before, we select six algorithms in four categories: random forest, Cascade Forest, GBDT, Xgboost, LightGBM, MLP, and TabNet, and the following is a brief introduction to each algorithm.

3.3.1. Random Forest. Random forest is an application of the bagging method (more precisely, the random patches algorithm) on decision trees. Specifically, the classic decision tree branching method is to divide all the attributes of the current node optimally. However, in random forest, it first randomly selects a certain number of attributes from the set of whole attributes of the current node as a candidate attribute subset and then selects the optimal division attributes from this candidate attribute subset. The scale of the attribute subset represents the degree of randomness, and the smaller the scale is, the stronger the randomness is. However, the stronger the randomness does not mean that the generalization ability of the random forest is much stronger. There is an optimal interval for the scale of the attribute subset. Random forest is the most typical representative bagging method and is the only algorithm of the bagging method that is comparative with the boosting method.

3.3.2. Cascade Forest. Cascade Forest is the core classification algorithm of deep forest, and its main idea is to learn from the neural network stack model. The input features or the features processed by multigranularity scanning constitute the first-level forest of Cascade Forest. The final output of each Cascade Forest is the classification results. At the next layer, those results are stacked with transformed input features. The final output is the average of all the previous outputs. In the Cascade Forest model, the output of the last layer is determined by two factors: the depth of the forest and the early stop. Therefore, Cascade Forest is a neural network whose neurons are replaced by decision trees and then added with residual connections.

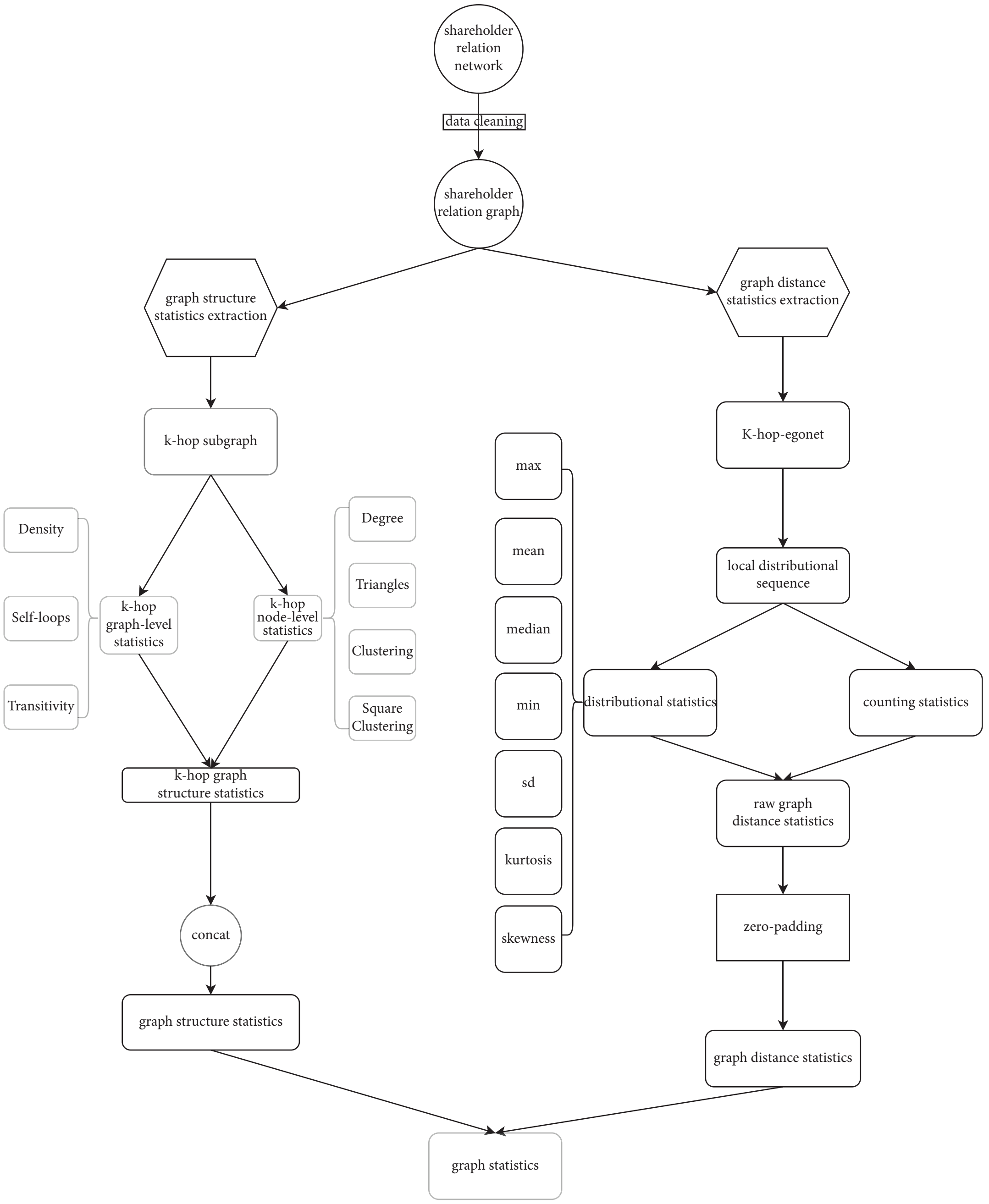
shareholder relation network
data cleaning
shareholder relation graph
graph structure statistics extraction
graph distance statistics extraction
k-hop subgraph
K-hop-egonet
Density
Self-loops
Transitivity
k-hop graph-level statistics
k-hop node-level statistics
Degree
Triangles
Clustering
Square Clustering
max
mean
median
min
sd
kurtosis
skewness
local distributional sequence
distributional statistics
counting statistics
k-hop graph structure statistics
raw graph distance statistics
zero-padding
concat
graph structure statistics
graph distance statistics
graph statistics

Figure 2: Extraction schema.

3.3.3. GBDT. GBDT is a breakthrough in early boosting methods. Unlike previous algorithms, GBDT does not improve model performance by modifying sample weights and learner weights but by fitting the residues of a previous learner with a new learner. Thus, the decision tree of GBDT is not a classification tree but a regression tree. The tree structure of GBDT is not used for classification but for regression of residuals, which belongs to the category of numerical prediction. Its optimization method is no longer in the parameter space based on analytical derivation but in the function space based on gradient descent.

3.3.4. Xgboost. Xgboost is an improved GBDT, whose improvements over GBDT include the following: (1) the single learner of GBDT only includes classification and regression trees, while Xgboost supports not only CART but also linear classifiers such as logistic regression; (2) Tte objective function of GBDT only includes classification/regression loss, while Xgboost adds a regular term on this basis to control the scale of the model and prevent overfitting; (3) GBDT uses the first order derivative based on the analytical formula in the optimization, while the Xgboost optimization is based on the second-order derivative; (4) GBDT follows the standard decision tree generation process in the node splitting process, which is very complicated when there are many feature values; at the same time, Xgboost uses a heuristic algorithm based on the decision tree, which selects several most promising nodes based on quantiles and divides the node selection from these nodes; (5) the learning rate of GBDT is constant during the whole training process, while Xgboost uses an adaptive shrinkage rate combined with leaf node weights; (6) GBDT does not support parallel computing, while Xgboost supports the process of node splitting at the same layer; and (7) Xgboost draws on the idea of a random forest for feature sampling, while GBDT does not.

3.3.5. LightGBM. LightGBM is an improved GBDT developed by Microsoft. It implements sample selection by GOSS and feature selection by EFB, which achieves a balance between accuracy and efficiency. Specifically, LightGBM has the following advantages compared with GBDT: (1) LightGBM implements a decision tree splitting strategy based on the histogram algorithm. It first converts continuous features into discrete features, then constructs a histogram based on those discrete features, and finally selects a splitting scheme based on the discrete eigenvalues. (2) LightGBM adapts to a leaf-wise decision tree growth strategy with additional depth constraints. The standard decision tree growth strategy is level-wise. More precisely, the leaf-wise strategy divides the leaf nodes of the same layer with equal weights, which brings unnecessary computational loss when the information gain of leaf node splitting is small.

In contrast, the leaf-wise strategy determines whether to split according to the information gained from the leaf nodes, which greatly reduces unnecessary calculations. (3) LightGBM directly supports categorical features. In GBDT and Xgboost, the algorithm cannot directly deal with the category features. More precisely, the category features need to be encoded in GBDT and Xgboost, which can easily generate an unbalanced decision tree. The space and time complexity will be high enough to express the full feature fully. At the same time, LightGBM divides the eigenvalues into two subsets, encodes them in binary groups, and then uses the histogram algorithm to optimize the regression tree, which greatly simplifies storage requirements and computational complexity. (4) LightGBM supports more powerful parallel computing. LightGBM can perform sample segmentation in parallel when constructing a local histogram based on the histogram algorithm, improving computational efficiency.

3.3.6. MLP. The neural network starts by simulating the feedback regulation mechanism of neurons. The basic structure of the neural network is the MP model, which accepts the input features by an input layer, weights them, then computes the summation, and finally outputs the result by an activation function. A typical and complete neural network consists of an input layer, a hidden layer, and an output layer. The number and the way of connection of the hidden layers determine the complexity of the neural network. The most typical neural network model is a multilayer feed forward neural network (MLP), in which each layer of neurons is completely interconnected with the next layer. However, there is no connection between the same layer and across layers.

3.3.7. TabNet. TabNet is a technology that uses deep neural networks to simulate the decision-making process of decision trees to process tabular data. It retains the advantages of end-to-end representation learning of DNN but also shares the advantage of interpretability and sparse feature selection of tree models. Such an advantage makes it comparable to the current mainstream tree-based ensemble learning models on tabular data. The complete unit of TabNet is a module that simulates a decision tree. The main difference between this module and MLP is the addition of a mask layer. The mask layer corresponds to the feature selection in the decision tree, while the fully connected layer plus ReLU corresponds to the threshold. Then, TabNet adds up the results of all condition judgments and gets the final output through the softmax layer. Typically, a TabNet unit contains four types of network layers: feature transformer, split, attentive transformer, and mask. In addition to the units, TabNet also contains a global feature attribute module. The Feature Transformer layer is used for the initial feature calculation, which consists of two parts, and the parameters of the first part are shared while the second part is not. The split layer cuts the vector output of the feature transformer layer into two parts, one part is used to compute the final output of the model and the other is used to calculate the next mask layer. The role of the attentive transformer layer is to calculate the mask layer of the current step based on the results of the previous step, whose main method is sparsemax calculation. The attention weights output by the attentive transformer layer differ for different samples. The global feature attribute module outputs the global importance of features.

4. Experiments

4.1. Experimental Settings. The main experimental steps and settings are described as follows: (1) align the data from different sources based on their unique IDs, generate 89 financial indicators based on expert experiences, clean the generated indicators, and delete invalid data. Then, use the standard normalization operation (standard scale) for continuous indicators and one-hot coding for discrete indicators to standardize them. Take the maximum and mean values of the historical returns for 2020 to construct the regression label. Take the ground truth of the default event to construct the classification label. The return rate of the default sample is set to 50% to expand the regression label. (2) For the shareholder relation network, use the statistics extraction schema depicted in Section 3.2 to get the full graph statistics vector. (3) Merge financial indicators and graph statistics vectors to get different feature combinations. (4) The sample set is divided into the training set and test set by 4 : 1. The training set is trained and verified by 4-fold cross-validation. (5) To evaluate the statistics and prediction models, we use MSE (mean squared error) for the regression task and recall for the classification task due to the proportion of positive samples being only 1%.

The main parameters of the machine are listed as follows: GPU: Nvidia RTX 2080Ti, CPU: Intel (R) Xeon (R) CPU E5-1630 v4 @ 3.70 GHz, RAM: 64 GB. The data source could be accessed in Gildata (see https://www.gildata.com/ for details), and the data we used contain about 944 samples for the regression task and 3052 samples for the classification task. There are 89 financial features, 9 graph distance statistics, and 21 graph structure features. For the consideration of commercial confidentiality requirements, we have to omit data descriptions. Interested readers can obtain the whole data from the official data source.

4.2. Experimental Results. The performances of features and models in the historical average return regression task are listed in Table 1, the performances in the historical maximum return regression task are listed in Table 2, and the performances in the default event classification task are listed in Table 3.

4.3. The Absolute Effectiveness of Graph Statistics of Shareholder Relation Network. As can be seen from Table 1, for the historical average return regression task, given the prediction model, the graph structure statistics achieve the best performance under two prediction models (MLP and TabNet). In contrast, distance statistics, structure statistics + distance statistics do not achieve the best performance under any prediction model. However, MLP and TabNet perform the worst among all the prediction models. Therefore, the graph statistics of shareholder relationship networks cannot be independently used as input features for predicting bond default risk. However, after combining with financial features, the graph structure statistics achieve the best performance under two prediction models (Cascade Forest and Xgboost), and the graph distance statistics achieve the best performance under one prediction model (GBDT). Graph structure statistics + graph distance statistics achieve the best performance under two prediction models (LightGBM and random forest). The combination of financial features + graph structural statistics + Cascade Forest achieves the second-best global performance, indicating that the graph statistics of the shareholder relation network can provide incremental information for predicting bond default risk based on financial features.

As can be seen from Table 2, for the historical maximum return regression task, given the prediction model, the graph structure statistics achieve the best performance under one prediction model (Xgboost), while graph distance statistics and graph structure statistics + graph distance statistics do not achieve the best performance under any prediction model, so the graph statistics of the shareholder relation network are not able to be independently used as input features for predicting bond default risk. However, after combining with financial features, the graph distance statistics achieve the best performance under two prediction models (random forest and TabNet), and the graph distance statistics + graph structure statistics achieve the best performance under three prediction models (GBDT, LightGBM, and MLP), indicating that the graph statistics of the shareholder relation network can provide incremental information for predicting bond default risk based on financial features.

As can be seen from Table 3, for the classification task of bond default event, the graph structure statistics achieve the best performance under three prediction models (GBDT, LightGBM, and random forest), and the graph distance statistics achieve the best performance under one prediction model (random forest). The graph statistics + graph distance statistics achieve the best performance under three prediction models (GBDT, random forest, and Xgboost). Therefore, graph statistics of shareholder relationship networks can be independently used as the input features of the classification task. After combining with financial features, graph structure statistics achieve the best performance under four prediction models (Cascade Forest, LightGBM, MLP, and random forest), and the graph distance statistics achieve the best performance under four prediction models (Cascade Forest, LightGBM, random forest, and TabNet). The graph structure statistics + graph distance statistics achieve the best performance under three prediction models (Cascade Forest, LightGBM, and random forest). Therefore, for classification tasks, the graph statistics of the shareholder relation network can provide incremental information based on financial features. However, because the labels of the classification task are too sparse, the reliability of the effectiveness of graph statistics on the classification task is insufficient.

In summary, graph statistics of shareholder relation networks could not be independently used as input features for predicting bond default risk. However, it can provide incremental information-based on their financial features.

4.4. The Relative Effectiveness of Graph Statistics of Shareholder Relation Network. As can be seen from Table 1, for the historical average return regression task, among graph structure statistics, graph distance statistics, and graph structure statistics + graph distance statistics, when used as independent input features, graph structure statistics perform better, which achieve the best performance under two prediction models (MLP and TabNet); when combined with financial features, graph structure statistics achieve the best performance under the two prediction models (Cascade Forest and Xgboost), and structural statistics + - distance statistics achieve the best performance under two prediction models (LightGBM and Random Forest). With the combination of structural statistics + financial features + cascade forest, it achieves global sub-optimal performance. Therefore, graph structure statistics is relatively the most effective for the historical average return regression task.

As can be seen from Table 2, for the historical maximum return regression task, among graph structure statistics, graph distance statistics, and graph structure statistics + - graph distance statistics, when used as independent input features, graph structure statistics perform better, which achieves the best performance under one prediction model

Table 1: The performances of features and models in the historical average return regression task.

	CF	GBDT	LGBM	MLP	RF	TabNet	Xgboost
fin	75.62	97.24	76.30	100.46	81.49	105.28	103.48
str	99.04	122.98	94.23	**92.91**	110.52	**95.08**	112.30
dis	94.49	113.34	94.96	95.74	109.31	111.40	113.41
str+dis	97.80	116.75	94.26	93.97	108.28	96.91	112.51
fin+str	**72.46**	98.45	72.02	99.46	80.08	97.46	**99.08**
fin+dis	72.60	**92.14**	74.85	94.70	81.40	101.79	103.48
fin+str+dis	75.53	93.39	**71.93**	97.33	**78.16**	96.71	100.64

Table 2: The performances of features and models in the historical maximum return regression task.

	CF	GBDT	LGBM	MLP	RF	TabNet	Xgboost
fin	***139.94***	162.53	155.17	216.03	145.08	163.79	179.63
str	168.46	203.02	166.18	*164.04*	207.83	179.19	176.35
dis	169.24	180.47	*167.76*	170.57	191.73	180.47	188.67
str + dis	166.73	193.12	*163.78*	170.96	203.74	175.82	187.19
fin + str	*141.30*	160.03	152.38	191.16	150.67	164.18	206.46
fin + dis	*141.12*	161.48	153.10	173.42	144.87	161.96	201.78
fin + str + dis	*143.75*	157.88	151.39	158.72	154.13	167.37	198.96

Table 3: The performances of features and models in the default event classification task.

	CF	GBDT	LGBM	MLP	RF	TabNet	Xgboost
fin	**1.0**	0.96	**1.0**	0.90	**1.0**	0.0	0.97
str	0.96	**1.0**	**1.0**	0.0	**1.0**	0.08	0.97
dis	0.95	0.89	0.97	0.0	**1.0**	0.06	0.94
str+dis	0.96	**1.0**	0.97	0.0	**1.0**	0.06	**1.0**
fin+str	**1.0**	0.94	**1.0**	**0.92**	**1.0**	0.0	0.97
fin+dis	**1.0**	0.94	**1.0**	0.48	**1.0**	**1.0**	0.97
fin+str+dis	**1.0**	0.98	**1.0**	0.13	**1.0**	0.14	0.97

(Xgboost); when combined with financial features, graph structure statistics achieve the best performance under two prediction models (random forest and TabNet), and graph structure statistics + graph distance statistics achieve the best performance under three prediction models (GBDT, LightGBM, and MLP). The combination of graph structure statistics + financial features + random forest achieves the optimal global performance. Therefore, graph structure statistics are relatively the most effective for the historical maximum return regression task.

As can be seen from Table 3, for the classification task, among graph structure statistics, graph distance statistics, and graph structure statistics + graph distance statistics, when used as independent input features, the structure statistics, structure statistics + distance statistics are comparable: each of them achieves the best performance under three prediction models. When combined with financial features, they are also comparable. However, again, because the labels of the classification task are too sparse, the reliability of the effectiveness of the graph statistics on the classification task is insufficient.

To sum up, among graph structure statistics, graph distance statistics, and graph structure statistics + graph distance statistics, graph structure statistics is relatively the most effective. Graph distance statistics are the least effective.

4.5. The Effectiveness of Prediction Models. As can be seen from Table 1, for the historical average rate of return regression task, given the input features, Cascade Forest achieves the best performance under three kinds of features (financial features, graph distance statistics, and financial + graph distance statistics). Lightgbm achieves the best performance under one kind of feature (financial features). MLP achieves the best performance under two features (graph structure statistics and financial feature + graph distance statistics). However, since the performance of MLP is far worse than that of LightGBM and Cascade Forest, LightGBM + financial features + graph structure statistics achieve the optimal global performance. Additionally, the number of optimal performance of features under Cascade Forest is more. Therefore, the best performance prediction model is Cascade Forest and LightGBM.

As can be seen from Table 2, for the historical maximum return regression task, given the input features, Cascade Forest achieves the best performance under three kinds of features (financial features, graph structure statistics + graph distance statistics, financial features + graph structure statistics, and financial features + graph distance statistics). LightGBM achieves the best performance on two features (graph distance statistics and graph structure statistics + - graph distance statistics), and MLP achieves the best performance on two features (financial features + graph structure statistics + graph distance statistics). Therefore, the best-performing prediction model is Cascade Forest, and the second best is LightGBM.

As can be seen from Table 3, for the classification task, Random Forest performs the best, which achieves the best performance under all input features. Cascade Forest and LightGBM are the second best, both of which achieve the best performance under four and more kinds of features. However, because the labels of the classification task are too sparse, the reliability of the effectiveness of the prediction models on the classification task is insufficient.

To sum up, for the historical average return regression task, Cascade Forest and LightGBM perform best for the historical maximum return regression task and Cascade Forest and LightGBM perform best. For classification tasks, random forest performs best.

5. Conclusions

In recent years, bond default events have occurred frequently in the China bond market, and bond default risk has risen sharply. However, traditional bond default risk prediction features only use limited transaction data and financial features, thereby being insufficient to predict effectively. Based on the theoretical starting point that the shareholder relationship network contains information about corporate governance structure and shareholders' willingness to rescue, we propose two kinds of graph statistics for the combination of the shareholder relation network and use Cascade Forest, GBDT, LightGBM, MLP, Random Forest, TabNet, and Xgboost as prediction models to test the absolute and relative effectiveness of the two kinds of graph statistics, the effectiveness of prediction models, and information of the shareholder relationship network.

Results show that for the historical average return regression task, given the prediction model, the graph statistics of the shareholder relation network cannot be independently used as input features for predicting bond default risk. However, it can provide incremental information based on their financial features. Among all kinds of graph statistics, graph structure statistics are relatively the most effective. Graph distance statistics are relatively the least effective. Given input features, Cascade Forest and LightGBM perform the best. Information about shareholder relation networks is effective. For the historical maximum return regression task, given the prediction model, the graph statistics of the shareholder relation network could not be independently used as input features for predicting bond default risk. However, it can provide incremental information based on their financial features. Among all kinds of graph statistics, graph structure statistics are relatively the most effective. Given the input features, Cascade Forest and LightGBM perform the best. Information about shareholder relation networks is effective. For the classification task, given the prediction model, graph statistics of the shareholder relation network not only can be independently used as the input features of the bond default classification task but also can provide incremental information based on financial features. The performances of all kinds of graph statistics are almost the same, which provides little incremental information. Given the input features, random forest performs best. However, because the labels of the classification task are too sparse, the reliability of the effectiveness of the shareholder relation network on the classification task is insufficient.

In summary, the graph statistics of shareholder relationship networks cannot be independently used as input features for predicting bond default risk. However, it can provide incremental information based on their financial features. Information about shareholder relation networks is effective for predicting bond default risk. Graph structure statistics perform the best among all features, and Cascade Forest and LightGBM perform the best among all prediction models.

Data Availability

All the Data that this article used are commercial, and are available at https://www.gildata.com/

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] W. Y. Loh, "Classification and regression trees," *WIREs Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[2] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[5] T. Chen and C. Guestrin, "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, New York, NY, 2016.

[6] G. Ke et al., "Lightgbm: a highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 30, pp. 3146–3154, 2017.

[7] Z. H. Zhou and J. Feng, "Deep forest," *National Science Review*, vol. 6, no. 1, pp. 74–86, Jan. 2019.

[8] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science*, vol. 14, no. 2, pp. 241–258, 2020.

[9] N. Gantchev and I. Chakraborty, "Does shareholder coordination matter? Evidence from private placements," *Journal of Financial Economics*, vol. 108, no. 1, pp. 213–230, 2013.

[10] T. H. D. King and M. M. Wen, "Shareholder governance, bondholder governance, and managerial risk-taking," *Journal of Banking & Finance*, vol. 35, no. 3, pp. 512–531, 2011.

[11] L. Garlappi, T. Shu, and H. Yan, "Default risk, shareholder advantage, and stock returns," *Review of Financial Studies*, vol. 21, no. 6, pp. 2743–2778, 2008.

[12] Y. D. Shi, M. Y. Song, F. Y. Li, and H. X. Zhen, "Share pledge of controlling shareholders and protection of corporate creditors interests: evidence from China's bond market," *Economic Research Journal*, vol. 56, no. 8, pp. 109–126, 2021.

[13] Z. C. Wu, Y. N. Zeng, H. Y. Zhou, and Y. C. Xu, "Mechanism of controlling shareholder's equity pledge on bond credit spreads," *Finance Research*, vol. 45, no. 5, pp. 78–88, 2021.

[14] M. Yıldırım, F. Y. Okay, and S. Özdemir, "Big data analytics for default prediction using graph theory," *Expert Systems with Applications*, vol. 176, Article ID 114840, 2021.

[15] J. W. Lee, W. K. Lee, and S. Y. Sohn, "Graph convolutional network-based credit default prediction utilizing three types of virtual distances among borrowers," *Expert Systems with Applications*, vol. 168, Article ID 114411, 2021.

[16] S. Lu, Y. Wang, X. Liu, and C. Jiang, "Multi-layer and parallel-connected graph convolutional networks for detecting debt

default in P2P networks," *Emerging Markets Finance and Trade*, vol. 58, no. 6, pp. 1688–1701, 2021.

[17] J. Lv, Y. Wang, and P. Guo, "Guarantee network risk contagion mechanism: path analysis and empirical test," *Management Review*, vol. 33, no. 3, pp. 1–13, 2021.

[18] B. Q. Li and X. Y. Zhang, "Research on the internal risk contagion mechanism of enterprise groups based on network structures," *Chinese Journal of Management Science*, vol. 29, no. 3, pp. 1–12, 2021.

[19] Q. Qian and K. Xu, "Research on the influence of entrepreneur's social relations on associated credit risk contagion—an dual-layer network perspective," *Chinese Journal of Management Science*, vol. 28, no. 11, pp. 35–42, 2020.

[20] X. Y. Wang and Z. Y. Zhang, "GR-LDA model with graph structure and its application in credit default warning," *Statistical Research*, vol. 38, no. 7, pp. 140–152, 2021.

[21] J. L. Zhang and J. Li, "Contagion and governance of enterprise group credit risk: an analysis based on complex network theory," *China Soft Science*, vol. 358, no. 10, pp. 119–136, 2020.

[22] C. Ying, T. Cai, S. Luo et al., "Do transformers really perform badly for graph representation?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 28877–28888, 2021.

[23] J. You, R. Ying, and J. Leskovec, "Position-aware graph neural networks," in *Proceedings of the International conference on machine learning*, pp. 7134–7143, Baltimore, Maryland, USA, 2019.

[24] P. Li, Y. Wang, H. Wang, and J. Leskovec, "Distance encoding: design provably more powerful neural networks for graph representation learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4465–4478, 2020.

[25] E. Alsentzer, S. Finlayson, M. Li, and M. Zitnik, "Subgraph neural networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8017–8029, 2020.