2023 IEEE International Conference on Data Mining Workshops (ICDMW) | 979-8-3503-8164-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICDMW60847.2023.00178

Industry credit spread forecasting based on time series data and industry relationships

Fan Meng
Department of Information Management,
Peking University,
Beijing, China
mengfan@pku.edu.cn

Yuyong Li
Department of Information Management,
Peking University,
Beijing, China
liyuyong@pku.edu.cn

Hanqing Yu
Department of Information Management,
Peking University,
Beijing, China
2100016634@stu.pku.edu.cn

Xiaofei Zhou
School of Cyber Security,
University of Chinese Academy of Sciences,
Beijing, China
zhouxiaofei@iie.ac.cn

Win-Bin Huang
Department of Information Management,
Peking University,
Beijing, China
huangwb@pku.edu.cn

***Abstract*—Credit spread is an important index in the financial market, as it captures the risk for the bond issuer and industry. As there are several types of connections among different industries, through which the risk may propagate from one industry to another, it may be helpful to take these connections into account for bond spread forecasting. In this paper, we propose a novel industry bond spread forecasting method based on LSTM and the temporal graph neural networks, which exploits the relations among different industries. The experimental results demonstrate that introducing industry relations can improve the prediction accuracy of the prediction.**

Keywords*—credit spread forecasting, temporal graph neural networks, LSTM

I. INOTRODUCTION

The credit spread is an important feature of credit bonds in the bond market, which measures the difference of interest rate between the bonds. The credit spread has been studied from many views[1-3] for macroeconomic analysis. However, compared to the stock price prediction[4-6], credit spread forecasting is not well studied. In China's bond market, the bond spread refers to credit spread, and the bond is generally divided into two kinds, which are interest rate bonds and credit bonds. The interest rate bond issuer is endorsement by country or by the credit of the central government, and it can be considered to have no credit risk. The issuer of credit debt often does not have the credit of the country, leading to a certain credit risk, so there is a credit risk compensation interest rate difference between the credit debt and interest rate debt, called credit spread.

In the bond market, the credit spread of a single bond is often affected by the credit rating and operating status of the issuer. The existing methods for predicting the credit spread of a single bond are often based on the historical credit and credit status of the bond as well as the characteristics of the individual bond[7-8]. Attention mechanism is also introduced to improve the prediction accuracy[9]. However, the bond spread index of the whole industry often lacks effective forecasting methods at present. The bond spread index of an industry has various and rich applications. On the one hand, the bond spread index of an industry can know that investors choose the industry with good prospects to invest in the bond market. On the other hand, the bond spread index of an industry can also detect the change of the direct credit risk of the industry from the perspective of the market.

We note that the bond spread index of the industry is often more cyclical, which is suitable for using the method of time series analysis to forecast. At the same time, we noticed that there are rich relationships between industries such as upstream and downstream, which may make the interest rate dynamics of corresponding industries have a certain co-movement effect. For example, the change of the bond credit spread in the upstream industry will affect the credit spread index of the downstream industry. Recently, many studies introduce graph neural networks to capture the relation among the data samples in a variety of prediction applications[10-13]. In[14], the authors combine weakly labeled learning method and temporal graph convolutional and make an accurate prediction of bond default prediction, and connections among industries has proven valuable in bond market prediction.

Based on the above observation, this paper aims to make reasonable use of both time series data, and relationships between industries for credit spread prediction. Specifically, upstream and downstream relationship of the industry, the historical time series similarity of bond spreads and historical similarity of stock price index are taken into account. The main purpose of this paper is to make a reasonable prediction of the bond spread of the industry, and explore the possible linkage effect in the process of dynamic change of the bond credit spread of the industry.

II. RESEARCH METHODS

Inspired by [15], which uses graph neural network to make rank prediction for stock, this paper introduces graph neural network method to integrate the historical data of bond spread and the relationships among industries for credit spread prediction. At the same time, we introduce attention to the ranking among bonds to the prediction target, which is more in line with the real investment scenario. The following part will introduce the research methods of this research in details.

A. Predicting the Target

When forecasting the bond spread of the industry, this paper absorbed the method of adding the ranking of the industry spread into the forecast target, which is more suitable for the real investment scenario.

2375-9259/23/$31.00 ©2023 IEEE
DOI 10.1109/ICDMW60847.2023.00178

TABLE I. EFFICIENCY OF RANK PREDICTION

Ground Truth			Method 1					Method 2				
			Prediction			Performance		Prediction			Performance	
A	B	C	A	B	C	MSE	Profit	A	B	C	MSE	Profit
35	15	-40	45	-5	-60	300	35	20	30	-40	150	15

As shown in the Table 1, if we only focus on the difference between the predicted value and the true value of different industries by different models (the index often used is MSE), and do not focus on the difference between the ranking relationship between the predicted value of different

industries by the model and the real ranking relationship, then according to the figure above, We will choose model 2 with a smaller MSE as the reference model to guide our investment. However, we can find that the ranking of the predicted values in model 2 is inconsistent with the real situation, while the ranking of the predicted values in Model 1 is completely consistent with the real situation. As a result, when we use model 2 to guide investment, we cannot select the industry with the highest possible return. Instead, it is a better choice to choose model 1, which has larger MSE but the ranking of predicted industries is more consistent with the real situation.

At the same time, we also consider the investment strategy of risk diversification in the real situation, so we assume that investors will adopt the investment strategy of risk diversification, that is, they will choose the industry with the best forecast in each industry to invest in different industries. Therefore, our prediction goal should focus on the accuracy of the ranking prediction among each industry within the same industry rather than the ranking prediction accuracy of all industries. The final objective function of our model is as follows:

$$l(R^t, r^t) = \left||R^t - r^t|\right|^2 + a \sum_{i=0}^{N} \sum_{j=0}^{N} \max(0, -(R_i^t - R_j^t)(r_i^t - r_j^t) a_{ijz}) \quad (1)$$

where R_i^t represents the true value of the industrial bond spread of industry i on the t day, r_i^t represents the predicted value of the industrial bond spread of industry i on the t day, a_{ijz} represents the z component in the relationship vector between industry i and industry j, which represents whether industry i and industry j belong to the same industry, if they belong to the same industry, the value is 1, otherwise the value is 0. In our objective function, the former term measures the gap between the predicted value of the model and the real value, that is, MSE, while the latter term measures the error between the ranking of the bond spread between the industries of the same industry and the real industry spread in the model prediction results, where a represents the parameter weighing the importance between the two terms.

B. *Time Series data processing -- LSTM sequence embedding layer*

First, we built an LSTM model to obtain the embedded expression of time series data. The structure of the model is shown in the Fig. 1. For the historical time series data of bond spreads of each industry, we input the historical spread data of the previous four days for the prediction of day t. And calculate the moving average of 5 days, 10 days, 15 days and 20 days for the historical spread data of each day and take the noise of the smooth data as the input of the LSTM model. In the LSTM, we set the unit size to 64. Finally, after the data of day t-1 is input into the model, The LSTM model will output a 64-dimensional hidden state vector h(t-1). We use the hidden state vectors of different industries to input into the fully connected neural network to predict the bond spread for each industry on day t. Therefore, we trained the LSTM model, and the hidden state vector h(t-1) output by each industry at time t of the model after training was expressed as the embedded vector e(t-1) of the historical bond spread data of the first four days of the t day.

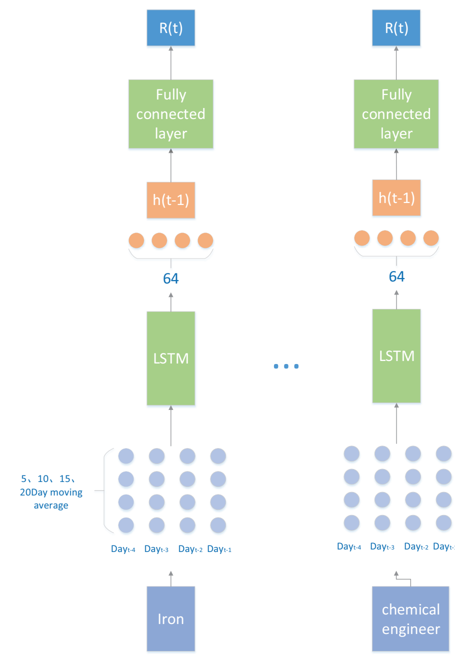

Fig. 1. LSTM sequence embedding layer

It needs to be pointed out and explained that the objective function of the LSTM model is basically the same as the objective function introduced in Section 3.1, except that we have not added the industry relationship here, so we do not consider the error of calculating the predicted ranking value by industry. Here we use this objective function to train the LSTM model, and its output hidden state vector h (T-1) can guarantee the effective prediction of the bond spread on the t day, so it can effectively express the characteristics of the historical bond spread data of the previous four days. This is the reason why h (t-1) is chosen as the embedded vector expression of historical bond spread time series data.

C. Relational data processing -- TGC diagram neural network layer

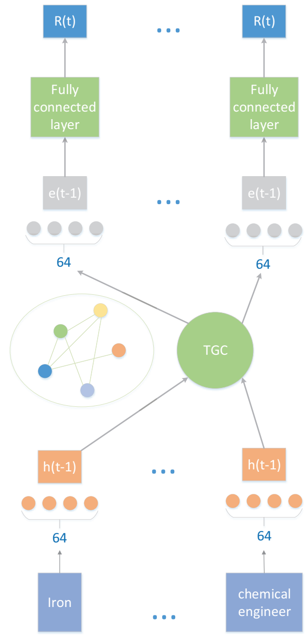

Fig. 2. TGC diagram neural network layer

After obtaining the embedded vector representation of the historical bond spread data, we need to fuse the industry-to-industry relational data with the series-embedded data. The traditional way of fusion is often to represent the relationship between an industry and other industries as a high-dimensional vector m(t-1), and then simply splicing h(t-1) and m(t-1) as inputs to a fully connected neural network, so as to make predictions. This fusion approach is often too crude and concatenates the relationship vector with the time series embedded vector by a preset weight, often missing a certain rationality.

Here we use the TGC graph neural network to embed the relational data. The core idea is that according to the relationship between industry i and industry j (represented by the relationship vector a_{ij}, which may have multiple dimensions, representing whether a variety of relationships between industry i and industry j exist or not, the corresponding dimension value is 1 if it exists, otherwise it is 0), the sequence embedding vector of industry i and industry j is updated with different weights. And then fuse the relational data into the embedding vector representation of the time series. The specific formula is as follows:

$$e_i^{t-1} = \sum_{\{j\,|sum(a_{ji})>0\}} \frac{g(a_{ji}, e_i^{t-1}, e_j^{t-1}, s_i^{t-1}, s_j^{t-1})}{d_j} e_j^t \qquad (2)$$

where e_i^t represents the embedded vector of the i industry at the prediction time t, $g(a_{ji}, e_i^{t-1}, e_j^{t-1}, s_i^{t-1}, s_j^{t-1})$ is when the j industry updates the i industry vector, the weight function of the weight is calculated ,according to the relationship between the two. The design of the weight function is the focus of the model. In the design of the weight function, we use the above mentioned: the relationship with the industry, the upstream and downstream relationship, the similarity relationship of historical bond spreads, and the similarity relationship of historical stock price index. The design of the weight function is as follows:

$$g(a_{ji}, e_i^{t-1}, e_j^{t-1}, s_i^{t-1}, s_j^{t-1}) = {s_i^{t-1}}^T s_j^{t-1} \times {e_i^{t-1}}^T e_j^{t-1} \times \emptyset(w^T a_{ji} + b) \qquad (3)$$

where the relation with the industry and the relation between upstream and downstream are expressed as one dimension of a_{ji}, w^T is the parameter content that our graph neural network needs to train, and according to w^T we can know the weight of the relation of each dimension. At the same time, this is also the content that we need to focus on later, which can reveal the linkage effect of the dynamic change of interest rate spread between different industries. ${s_i^{t-1}}^T s_j^{t-1} \times {e_i^{t-1}}^T e_j^{t-1}$ represent the similarity of the historical stock price index and the historical bond spread between the two industries, respectively, $e_j^{t-1} e_i^{t-1}$ are the embedding vectors of the historical bond spread of the industries i and j. We believe that if the similarity between the two embedding vectors is high, it means that the two industries have a certain similarity in the change of interest spread in the past time. Then the relationship between industry i and j should be closer, which will increase the value of the weight function. Similarly, the embedding vector of the stock price index of the industry ${s_i^t}^T s_j^t$ is also obtained according to the same LSTM layer training method. We believe that the stock price index of the industry reflects the current economic conditions of the industry to some extent, and the more similar the economic conditions of the two industries, the greater the impact of the spread between the two industries should be, thus increasing the value of the weight function. Therefore, in the TGC graph neural network layer, we use the above method to fuse the industry relationship data into the embedded data of the industry's historical bond spread, to complete the purpose of data fusion.

D. Prediction Layer

With the above foundation, the design of the prediction layer is relatively simple. Our graph neural network layer will finally output a 64-dimensional embedded vector e_i^{t-1} updated by graph neural network when predicting the spread of time t for each industry. When we predict time t, We can input the embedded vector of each industry into the fully connected neural network layer to predict the predicted value of the bond spread r_i^t of the final industry at time t.

III. EXPERIMENTAL DATA

When conducting specific experiments, it is necessary to collect time series data of bond spreads of different industries, as well as relationship data between industries.

Here, we choose the 28 industries under Shenwan's classification of first-level industries, excluding the remaining first-level industries after comprehensive and banking, and some second-level industries, a total of 36 industries are taken as the industry set that needs to be predicted in this experiment.

For the time series data of bond spread, the industrial bond credit spread data of the closing price of 2787 trading days from 2010-1-4 to 2021-3-24 of Shenwan's corresponding industries were obtained from the wind database.

For the relationship data between industries, we based on the existing 36 industries, first of all, according to the National Bureau of Statistics of China in 2017 three industrial classification and national economy industry classification data, the collected industries to build a classification system, the construction of industrial relations between industries, part of the structure is shown in Fig. 3.

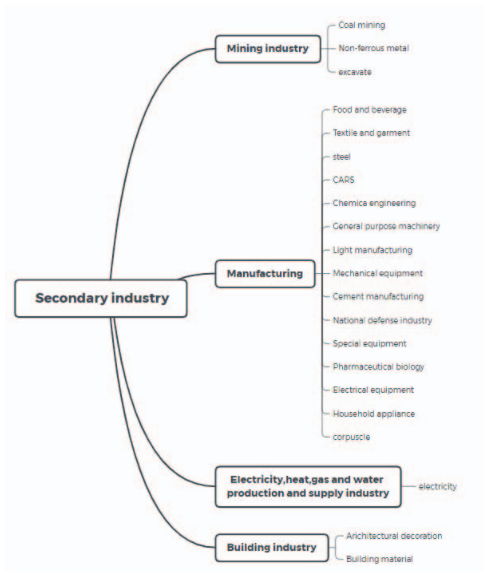

Fig. 3. Part of the industry structure

Then for the upstream and downstream relationship between industries, we extracted the direct upstream and downstream relationship between 57 industries from the upstream and downstream relationship chart in Fig. 4 and built the indirect upstream and downstream relationship based on this.

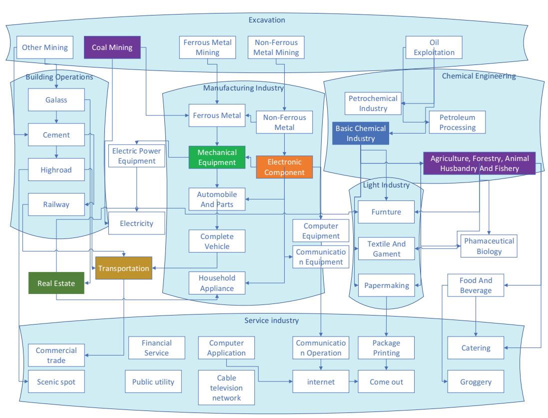

Fig. 4. Upstream and downstream relationship

For the historical bond spread similarity relationship, the time series data obtained above are used. For the historical stock price similarity relationship, corresponding to the 36 industries used in the bond spread data and the corresponding time interval (2010-1-4 to 2021-3-24), the industry stock index data of Shenwan was obtained.

When training and testing the model, we divided the data into the training set, verification set and test set according to the ratio of 3:1:1 according to the time interval.

IV. Experimental Results

During the experiment, we distinguish five different models, among which the most basic model is Rank_Lstm model, which only uses LSTM model and time series data to make prediction. For the model with graph neural network added, we make certain distinctions in the use of relationships. In the same industry relationship and upstream and downstream relationship established above, we have six relationships belonging to the same primary industry, the same secondary industry, the direct upstream, the indirect upstream, the direct downstream and the indirect downstream, among which all belong to the secondary industry. Located in the direct upstream, located in the direct downstream, these three relationships are defined as direct relationships, so we distinguish the model using three direct relationships and the model using all six relationships, respectively, marked 3rel and 6rel. In addition, in terms of the objective function of the model, on the one hand, we use the objective function model in the way of industry division, and on the other hand, we use the objective function model that does not distinguish industry and calculates the error between the predicted ranking values of all industries together, which is labeled as separate and all. Therefore, in the experiment, we compare the total of 5 different models.

A. Model prediction results

TABLE II. Model Prediction Result

Model	MSE	MRRT	BackTest
Rlstm_separate6rel	5.258 e-4	1.92 e-1	6.29
Rlstm_separate3rel	5.279 e-4	1.80 e-1	6.32
Rlstm__all6rel	5.259 e-4	1.86 e-1	5.62
Rlstm__all3rel	5.278 e-4	1.76 e-1	5.70
Rank_Lstm	5.306 e-4	1.82 e-1	3.79

In the above table, we use three criteria to measure the performance of the model. The first is MSE, which measures the accuracy of the predicted value, and the second is Mrrt, which measures the accuracy of the model's ranking of the predicted value among industries. Mrrt is a measurement index often used in information retrieval system to measure the accuracy of the hit results of the retrieval system. We pay attention to the accuracy of the real industry with the highest bond spread, if the industry ranks the NTH in the predicted results, then its score value is 1/n, and the Mrrt value is obtained by synthesizing all the predicted results. BackTest, on the other hand, is the result of an investment return based on the model, which will be described in detail in the next section.

Here we can clearly see that the Rlstm_separate6rel model performs best, its MSE and Mrrt level 1 Bt are the best models, while the Rank_Lstm model performs poorly, not only its MSE result is the highest, but also the Bt backtest result is the worst. Moreover, it can be clearly found that almost all models with added relational data and graph neural network layer perform far better than LSTM models using only time series data. Moreover, the models that use all relationships and only calculate the inter-industry ranking error are superior to other models. This shows that the indirect relationship is still useful, and that the way in which the interindustry ranking error is calculated based on the actual investment situation produces better predictions, especially in the backtest results.

B. Backtest results

As we want to test our model's ability in investment, we design the model's backtest strategy. Based on the strategy of diversifying risk investment, bonds with the highest and lowest predicted credit spread is selected from 10 secondary industries each day, regardless of market factors such as liquidity. Assuming that there is a derivative that can directly trade credit spread, the difference between the highest and lowest interest spread can represent the available returns. We sum up the results in time series. When comparing the models, compare the cumulative curve of the model with the cumulative curve (baseline) of the strategy that selected the median of the values of the true industry bond spread on a daily basis.

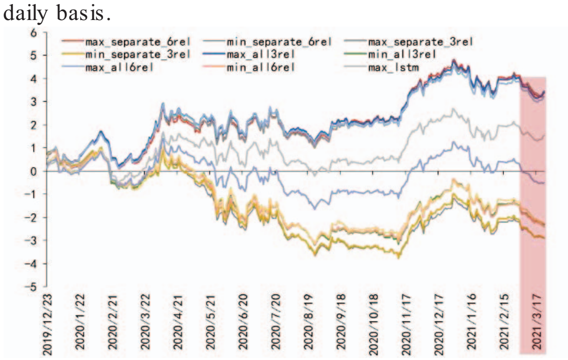

Fig. 5. the cumulative curve of the model

As can be seen from the results in the Fig. 5, basically all models performs better than the baseline in predicting the maximum spread and smaller than the baseline in predicting the minimum spread, which illustrates the validity of our model in an investment-based backtest strategy. We subtract the maximum value from the minimum value to get the possible return in the backtest investment strategy, and then compare the performance of different models. The results are shown in Fig. 6.

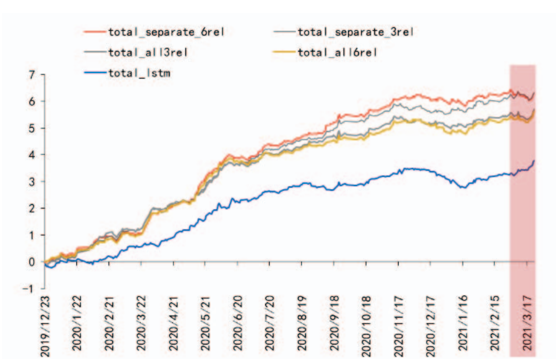

Fig. 6. Backtest results

From the backtest results, it is clear that the model using the proposed objective function to calculate the error of the predicted ranking values by industry is significantly better than the other models, and still the model using all the direct and indirect relationships performs better. In addition, the conclusion is the same as in the previous section, the model with industry relationship data and graph neural network layer is much more efficient than the base model using only LSTM model and time series data in the backtest.

C. Linkage effect of industrial interest rate spread dynamics

The conclusion above has effectively demonstrated that our model can effectively describe the linkage effect between the dynamics of industry interest rate differentials. Now we are concerned about the performance of this linkage effect. We look in the weight function w^T to see which relationships between industries have the stronger linkage effect. We select the best-performing model above -- and discuss the results of all six direct and indirect relationships using industry-specific calculations of the ranking error objective function. The weights of the relationships are shown in Fig. 7.

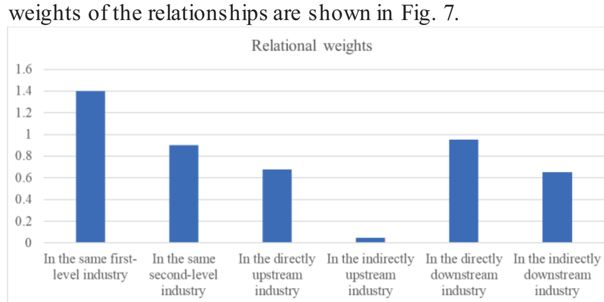

Fig. 7. Weight function

From the figure above, it is obvious that the dynamic linkage effect of the interest rate difference between the two industries that belong to the same primary industry relationship and the secondary industry relationship is large. This may be related to the policy direction, for example, after the introduction of relevant policies of supply-side reform, many manufacturing industries with excess capacity tend to suffer from both prosperity and loss. At the same time, it is also closely related to the shift of economic focus in the process of national development. For example, with the process of reform and opening up, the primary industry will be more and more marginalized. As a result, some agriculture, forestry, animal husbandry and fishery industries in the primary industry will often face deteriorating economic conditions at the same time. As a result, enterprises in the primary industry often have to pay a higher interest rate to obtain financing when issuing bonds, that is, the linkage effect of the bond spread of many industries in the primary industry increases at the same time.

At the same time, we can also find some phenomena worth thinking about. In the upstream and downstream relationship of the industry, the linkage effect of the direct upstream and downstream relationship is greater than the indirect upstream and downstream relationship. At the same time, the downstream industry has a greater linkage effect on the upstream industry, so why the downstream industry has a greater linkage effect on the upstream industry is an interesting phenomenon, which is contrary to our common sense assumptions. Here we consider a possible explanation, as today's Chinese economy is facing a serious overcapacity situation, the main constraints of economic growth in the demand side, and the downstream industry is located in the upstream industry demand side, the upstream industry is located in the downstream industry supply side, it is this situation makes the downstream industry differential dynamic can effectively feed back to the upstream industry, And the dynamic of the spread of the upstream industry is more difficult to feedback to the downstream industry. Here we consider a simple example:

The operation of some industry giants in the downstream industry has problems -- > The overall operating condition of the downstream industry has deteriorated -- > The overall demand of the downstream industry to obtain financing to improve its operating condition has increased -- > Enterprises in the downstream industry have generally increased the interest rate of bonds issued to obtain financing in the capital market -- > The bond spread index of the downstream industry has increased.

At the same time, we can find that some of the industry giants in the downstream industry have problems -- > The overall operating condition of the downstream industry is deteriorating -- > The downstream industry tends to reduce the scale of operation, the demand for the upstream industry is reduced -- > The upstream industry has overcapacity, the price is basically maintained at about the cost price, the demand is reduced and the price cannot rise (because of the fierce competition caused by overcapacity), Ultimately, the overall operating condition of the upstream industry deteriorates -- > Using the same logic as the downstream industry above: the bond spread index of the upstream industry rises.

In this way, we can obviously see that the rise of the bond spread index of the downstream industry will soon lead to the rise of the bond spread index of the upstream industry. Similar logic also applies to the decline of the bond spread index of the downstream industry.

Next, consider why the upstream industry has less impact on the downstream industry from another aspect. First, based on similar logic: Some industry giants in the upstream industry have problems -- >…-- > The bond spread index of the upstream industry rose.

The operation of some industry giants in the upstream industry has problems -- > The overall operating conditions of the upstream industry have deteriorated -- > The upstream industry has reduced some production capacity in order to improve the operating conditions - > Due to overcapacity, although the upstream industry has reduced some production capacity, the overall supply in the market is still in a surplus situation, and the fierce competition still leads to the price of products in the upstream industry approaching the cost price. Will not change significantly -- > The downstream industry is basically unaffected.

Through this logical reasoning, based on the current situation of excess capacity in our country, the impact of all upstream and downstream industries is integrated, so as to show the result as a whole: the downstream industry's linkage impact on the upstream industry is greater than the upstream industry's linkage impact on the downstream industry.

V. CONCLUSION

In this study, we use the historical bond spread time series data of the industry and the relationship between the industries to successfully complete the task of predicting the bond spread of the industry. At the same time, we also find that when the graph neural network model and the relationship data of the industry are introduced, the performance of the model is much better than that of the LSTM model which only use the time series data. On the one hand, this shows the effectiveness of our model; on the other hand, the excellent performance of the model also shows that we can effectively depict the linkage effect between the dynamics of the bond spread of the industry. We found that the linkage effect of the industry with the industry is larger, the linkage effect between the direct upstream and downstream industries is larger, and the linkage effect of the downstream industry on the upstream industry is greater.

In addition, there are several shortcomings in this study. We consider the following directions for improvement: further enrich the relationship between industries, use a graph model with added relational data for comparison, use models from other data fusion methods for comparison. To further explore the spread prediction of individual bonds, event data, emotional data, financial risk index data of issuing entities can be introduced in this task.

Acknowledgments

This work was supported by National Natural Science Foundation of China (#71801232, #71932008) and The Fundamental Research Funds for the Central Universities, Peking University.

References

[1]. De Santis, Roberto A., and Wouter Van der Veken. "Forecasting macroeconomic risk in real time: Great and Covid-19 Recessions." (2020).

[2]. Cicchiello, Antonella Francesca, et al. "Credit spreads in the European green bond market: A daily analysis of the COVID-19 pandemic impact." *Journal of International Financial Management & Accounting* 33.3 (2022): 383-411.

[3]. Lian, Yonghui, et al. "How does corporate ESG performance affect bond credit spreads: Empirical evidence from China." International Review of Economics & Finance 85 (2023): 352-371.

[4]. Lil, Angela, et al. "FLAG: Stock Movement Prediction via Fusing Logic and Semantic Graphs of Financial News." 2022 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2022.

[5]. Lu, Wenjie, et al. "A CNN-BiLSTM-AM method for stock price prediction." Neural Computing and Applications 33 (2021): 4741-4753.

[6]. Li, Xiaodong, Wu Pangjing, and Wang Wenpeng. "Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong." Information Processing & Management 57.5 (2020): 102212.

[7]. Wang, Chunjing , and J. Qu . "Analysis of the Pro-cyclical Behavior of Credit Spread in Chinese Bond Market." Journal of Applied Finance & Banking 10(2020).

[8]. Krishnan, C. N. V., Peter H. Ritchken, and James B. Thomson. "Predicting credit spreads." Journal of Financial Intermediation 19.4 (2010): 529-563.

[9]. Liu, Xiao, et al. "A Novel Methodology for Credit Spread Prediction: Depth-Gated Recurrent Neural

Network with Self-Attention Mechanism." Mathematical Problems in Engineering 2022 (2022).

[10]. Cheng, Dawei, et al. "Financial time series forecasting with multi-modality graph neural network." Pattern Recognition 121 (2022): 108218.

[11]. Yang, Yiying, et al. "Using external knowledge for financial event prediction based on graph neural networks." Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019.

[12]. Xiang, Sheng, et al. "Temporal and Heterogeneous Graph Neural Network for Financial Time Series Prediction." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.

[13]. Wang, Daixin, et al. "A semi-supervised graph attentive network for financial fraud detection." 2019 IEEE International Conference on Data Mining (ICDM). IEEE, 2019.

[14]. Li, Yuyong, et al. "Bond Default Prediction with Temporal Graph Convolutional Neural Network and Weakly Supervised Learning." Procedia Computer Science 221 (2023): 1376-1385.

[15]. Feng, Fuli, et al. "Temporal relational ranking for stock prediction." ACM Transactions on Information Systems (TOIS) 37.2 (2019): 1-30.