Postprint: High Default Risk User Identification Algorithm in Internet Finance Platforms

Authors: Yang Xiaohui, Guo Binghui, Mi Zhilong, Zheng Zhiming

Date: 2018-05-18T00:00:00+00:00

Abstract

In the context of continuous innovation in financial technology in China, employing network analysis techniques for user risk identification in internet financial platforms has become a key focus in current technological development. This study takes user transaction data from an internet financial platform as the research object, analyzes the propagation behavior of loan delinquency and default within the data, and proposes constructing model algorithms based on propagation characteristics to identify high-risk users on internet financial platforms. Building upon SIS and SIR models constructed with threshold-based and random propagation mechanisms, the models are transformed into algorithms capable of evaluating user risk values, which are further validated and compared against actual default data. Comparative results demonstrate that under the high-risk group segmentation criteria of top 5% and 10%, the algorithm achieves high recall rates and exhibits good structural correlation.

Full Text

Preamble

Identification Algorithm for High Breaching Risk Users in Internet Financial Platforms

Yang Xiaohui, Guo Binghui†, Mi Zhilong, Zheng Zhiming
(School of Mathematics and Systems Science, Beihang University, Key Laboratory of Mathematics, Information and Behavior of the Ministry of Education, Beijing 100191, China)

Abstract: Against the backdrop of continuous financial technology innovation in China, employing network analysis techniques for user risk identification on internet financial platforms has emerged as a prominent technical direction. This study examines user transaction data from an internet financial platform and proposes a model algorithm that identifies high-risk users by analyzing the

propagation characteristics of loan default behavior. Building upon SIS and SIR models based on threshold propagation and random propagation, we transform these models into algorithms capable of evaluating user risk values and validate them against actual default data. Comparative results demonstrate that the algorithm achieves high recall rates and strong structural correlations when identifying the top 5% and 10% high-risk groups.

Keywords: risk propagation; complex networks; internet finance; identification algorithm

0 Introduction

With China' s robust economic development and continuous financial technology innovation, internet financial platforms have rapidly expanded in the financial market. Against this backdrop of increasingly frequent personal lending, these platforms enable networked operations of loan services through internet channels, with P2P lending platforms being the most typical representatives. Xiong Yahua et al. conducted detailed investigations into the credit mechanisms, theoretical foundations, and risk sources of P2P lending [1]. To ensure that internet financial platforms healthily integrate into the existing financial system under regulatory policies, research on constructing systemic risk control models and conducting user risk level identification using data and network analysis technologies has become a cutting-edge topic with practical significance.

Scholars have previously approached default risk from various perspectives. Wang Shubin et al. [2] reviewed the current state and methodologies for assessing P2P lending default risk and its contagion. Wang Qian et al. [3] modeled and empirically analyzed credit default risk contagion patterns and regularities. Li Jieting [4] conducted modeling and simulation studies on correlated credit risk contagion using complex networks. Chen Tingqiang et al. [5] systematically analyzed the evolutionary mechanisms of credit risk contagion behavior on complex networks. Wang Shenkan et al. [6] employed information asymmetry theory and linear regression analysis to investigate the correlation between default rates and factors such as borrowing interest rates, loan terms, and credit ratings. Ding Lan et al. [7] constructed an ensemble strategy model based on primary and secondary learners to evaluate and predict user default risk in P2P lending. Tang Jianqin et al. [8] used an improved C4.5 decision tree model for credit assessment, yielding highly explanatory conclusions.

In internet financial platforms, where credit profiles are constructed from internet data rather than direct business contact, most user default risk predictions rely on association analysis between user credit risk and user characteristics. However, factors affecting user default that stem from network relationships remain underexplored. Based on analysis of actual transaction data and user network association structures, this paper investigates passive default user behavior on an internet financial intermediary platform, identifies debt default transmission characteristics, and constructs models using complex network con-

tagion diffusion models to identify high-risk groups vulnerable to passive default. By defining deterministic and threshold-based propagation patterns, we propose risk propagation models that reflect how defaults by superior nodes cause passive defaults, considering both default-immune and non-immune scenarios. The model quantifies systemic risk and high-risk group identification in internet financial platforms based on social networks. Using complex network structural parameters such as degree distribution, PageRank centrality, and betweenness centrality [9] as fundamental parameters for diffusion behavior, we obtain individual user default risk propagation evaluation values through network propagation simulation. Comparison with actual platform data reveals that the model effectively identifies high-risk groups in lending relationships who are susceptible to others' defaults and likely to cause defaults in others.

1 Risk Propagation-Based Identification Model

1.1 Model Framework

Analysis of lending data from an internet financial platform reveals that a considerable proportion of default transactions posing systemic risk result from defaults by superior nodes in the lending chain. This paper proposes a default risk propagation identification model based on default thresholds and propagation influence. The user default threshold is set according to the average value of all lending transactions by that user over the past 12 months. Since internet financial platforms typically restrict lending behavior for users who become delinquent, we propose the unrestricted QSIS model to describe systemic platform risk identification under unconstrained user lending behavior, and the restricted QSIR model for risk identification under constrained lending behavior. Users are categorized into non-defaulting (S), defaulting (I), and restricted trading (R) states. From a complex network perspective, considering network structure effects when multiple transactions occur simultaneously, we examine both deterministic and probabilistic propagation patterns under three structural parameters: degree distribution centrality, betweenness centrality, and PageRank centrality. The overall research framework is shown in [Figure 1: see original paper].

1.2 Deterministic Default Risk Propagation Model (QSIS)

We construct the network G(V, E) from actual transaction data, where V represents the set of all transaction users. In the QSIS model, users are divided into two node types: non-defaulting (S) and defaulting (I), with S(x) denoting the set of non-defaulting users and I(x) the set of defaulting users. E represents the set of transactions between users, with edge weights {a } representing transaction amounts. Since multiple transactions often exist between user pairs, we compress multi-edges and select the maximum transaction amount between any two users as their representative transaction amount for diffusion simulation.

Let n denote user i' s state, where n = 1 indicates default and n = 0 indicates

non-default. Let e denote the state of transaction E between users, where e = 1 indicates the transaction defaults and e = 0 indicates it does not. The amount user j is defaulted on is d = Σ A a e . If the threshold distribution for user default amounts is { }, we define the number of times a user is infected by propagated default as the user' s default propagation strength, reflecting the intensity of being "defaulted upon" in the lending network, denoted as {F }. Let L (x) be the set of "defaulted upon" users reached by default risk propagation throughout the diffusion process. The QSIS deterministic threshold propagation model can be described as follows [12]:

a) Initial state: All users are in non-default state (S). Randomly select a portion of users to become defaulting (I), i.e., randomly change some n from 0 to 1, causing a specific outstanding transaction E of these users to default, setting e = 1.

b) Defaulted transaction amounts accumulate onto creditors. Once a creditor' s defaulted amount exceeds their tolerance threshold, i.e., when d = Σ A a e > , the creditor' s state changes from S to I. This newly defaulted user will default on their specific outstanding transactions, with default amounts accumulating onto their creditors.

c) When all user states cease changing, one network propagation process completes. For the next propagation round, all users return to non-default state (S), regaining default potential. Randomly select a portion of users to default and begin the next propagation round.

d) Repeat the above simulation process, recording the number of times each user becomes infected and transitions to state I, which defines their default propagation strength. The propagation process terminates when the set of defaulted users L (x) remains unchanged between two consecutive propagation rounds.

1.3 Probabilistic Default Risk Propagation Model (PSIS)

The probabilistic PSIS model differs from the deterministic QSIS model in that when a user defaults, they no longer deterministically default on a specific transaction but instead default on one or several transactions with certain probabilities. In this paper, the specific steps differ from the QSIS model in the first and second stages: when user i defaults, transaction E will default with a probability related to the three network structural parameters mentioned above.

1.4 Restricted Models Based on Both Propagation Types (QSIR and PSIR)

Due to practical risk control requirements, internet financial platforms may restrict borrowing and lending transactions for users with high default risk. Under restricted trading conditions, we propose restricted deterministic and probabilistic default risk propagation processes, constructing corresponding identification

models QSIR and PSIR. The difference between QSIR and QSIS lies in the random immunization strategy applied during each propagation round, where a portion of users are randomly immunized, making them immune to defaults from superior nodes and unable to propagate defaults to inferior nodes.

2 Experimental Validation

2.1 Data Description

2.1.1 Original Dataset Data were obtained from a financial lending platform, selecting 3,312 users and all transaction records between May 14, 2015 and April 24, 2017, totaling 860,999 transactions. Among these, 872 users defaulted, with their sequence denoted as D(x). Analysis revealed that 213 users defaulted due to defaults by their debtors. These users' sequence is denoted as L(x). Multi-edges between users were compressed to construct the network ([Figure 4: see original paper]), from which default transactions were extracted to form a sub-network ([Figure 5: see original paper]).

2.1.2 Comparison Dataset Transaction data for the same users from July 22, 2016 to July 22, 2017 were selected as the comparison dataset. During this period, 3,312 users were involved in 537,146 transactions, with 11,657 default transactions. A total of 711 users defaulted on others, among which 43 were propagation defaults. This comparison dataset is denoted as C(x). Our models were applied to C(x) to calculate recall rates, denoted as R(x).

2.2 Network Structure Parameters

This study employs three network structure parameters:

a) **Degree centrality in directed networks**: In-degree k = Σ A and out-degree k = Σ A , where {A } is the adjacency matrix.

b) **PageRank centrality**: x = Σ A + , where and are positive constants.

c) **Betweenness centrality**: x = Σ n / g , where n is the number of geodesic paths from s to t passing through i, and g is the total number of geodesic paths from s to t.

2.3 Threshold Selection

Between May 14, 2015 and April 24, 2017, the 3,312 selected users conducted 860,999 transactions, including 16,457 default transactions. Within a 95% confidence interval, the amount distribution of all default transactions is shown in [Figure 6: see original paper]. The average transaction amount under the 95% confidence interval is 10,367. Therefore, this study selects 100 thresholds uniformly distributed from 100 to 10,000 for experiments.

2.4 Model Validation

We uniformly select 100 thresholds between 100 and 10,000 and conduct default risk diffusion studies under four models: QSIS, PSIS, QSIR, and PSIR. For each model, we analyze diffusion behavior differences under degree distribution, betweenness distribution, and PageRank distribution. Under each diffusion mode, we obtain the set of defaulted users L (x) and the default strength distribution {F } of these users. To compare prediction accuracy for defaulting users and users "defaulted upon" due to risk propagation across different diffusion models, we define recall rate as the ratio of correctly predicted actually-propagated defaulters to correctly predicted actual defaulters:

Recall = |L (x) L(x)| / |L (x) D(x)|

The recall rate reflects the predicted proportion of defaulting users who defaulted through propagation, indicating the accuracy of local diffusion default user predictions and serving as a valuable metric for evaluating propagation default prediction precision. We compare diffusion results for the top 10% and 5% high-risk groups across different models.

In PSIS and QSIS models, users are divided into non-defaulting (S) and defaulting (I) states based on the constructed transaction network. Each propagation randomly selects 600 users to default, observing default risk propagation in the network. In PSIR and QSIR models, users are divided into three categories: non-defaulting (S), defaulting (I), and blacklisted (R, also called restricted trading users). Similarly, each propagation randomly selects 600 users to default. When the set of "defaulted upon" users L (x) no longer changes during risk diffusion, the diffusion process converges and stops. We calculate the sequence of users in L (x) sorted by F and select the top 10% high-risk users (the top 332 users by influence {F }). The comparison results for QSIS and PSIS models under unrestricted lending are shown in and .

**** shows that under unrestricted lending, PSIS with degree distribution propagation achieves the highest accuracy in predicting propagation defaulters among defaulting users. Overall, QSIS demonstrates stronger prediction performance than PSIS in terms of predicting the proportion of propagation users among defaulters. However, in QSIS, PageRank-based predictions identify almost exclusively propagation defaulters—a conclusion that also holds for the comparison dataset.

**** presents PSIS model results for the top 10% high-risk users. **** and **** show PSIR and QSIR model results for the top 10% users, respectively. Comparison reveals that PSIR with degree distribution propagation achieves the highest accuracy in predicting propagation defaulters among defaulting users. Generally, SIR models with probabilistic propagation demonstrate higher accuracy for propagation defaulters than deterministic SIR models, though overall prediction accuracy is lower on the comparison dataset.

To further evaluate model performance on extremely high-risk groups, we select

the top 5% high-risk users (the top 166 users by influence {F }) and compare prediction accuracy under unrestricted lending. Results for QSIS and PSIS models are shown in and .

Comparison of through reveals that PSIS model prediction accuracy for propagation defaulters is higher in the top 10% high-risk group than in the top 5% group—a conclusion that also holds for QSIS (comparing and). This indicates that unrestricted lending models show decreased accuracy when predicting users with higher propagation risk. Notably, shows the same phenomenon as : PageRank-based predictions identify few high-risk users but with 100% precision (all are propagation defaulters), a conclusion that also holds for the comparison dataset.

For restricted lending models, we similarly select the top 5% high-risk users (top 166 by influence {F }) and compare prediction accuracy. QSIR and PSIR results are shown in and .

Analysis of the original dataset shows that PSIR model prediction accuracy for propagation defaulters is generally higher in the top 10% group than in the top 5% group, except for betweenness propagation. This conclusion also holds for QSIS (comparing and), indicating that restricted lending models show decreased accuracy when predicting higher-risk propagation defaulters, though betweenness propagation achieves higher precision for the most dangerous propagation defaulters.

Comparing both datasets reveals that overall prediction performance is worse on the comparison dataset, but PageRank demonstrates better concentrated prediction performance.

Overall analysis of through for the original dataset shows that models employing random immunization strategies (PSIR and QSIR) achieve better prediction accuracy for propagation defaulters in lending networks. For the original dataset, all four models demonstrate higher prediction accuracy for the top 10% high-risk group than for the top 5% group. Except for QSIS, betweenness centrality achieves better results than the other two structural parameters for predicting top 5% high-risk propagation users, while PageRank centrality diffusion performs better than the other parameters for top 10% high-risk propagation users (except in PSIS).

Comparing both datasets, the comparison dataset shows worse overall prediction performance, but PageRank' s concentrated prediction performance is superior. Since the comparison dataset covers one year while the original dataset spans nearly two years, this suggests that the models perform better on longer-term data. For shorter-term datasets, while overall prediction accuracy decreases, PageRank-based predictions identify high-risk users who are all propagation defaulters with 100% precision, and this occurs more frequently than in the original dataset, indicating that short-term datasets better predict the concentrated nature of propagation defaults.

2.5 Model Comparative Analysis

The preceding analysis focused on accuracy. We now examine the distribution of user propagation influence F under three structural parameters across four propagation modes (QSIS, QSIR, PSIS, PSIR). Selecting a threshold of 5,000, we normalize and plot the distributions. Original dataset results are shown in [FIGURE:7.1], [FIGURE:8.1], and [FIGURE:9.1]; comparison dataset results appear in [FIGURE:7.2], [FIGURE:8.2], and [FIGURE:9.2].

Comparison across the three network structural parameters reveals that under the same parameter, user default propagation strength distributions differ little across diffusion modes, generally following exponential distributions. However, PSIS model distributions under betweenness and PageRank propagation show significant differences from other models, with influence distributions approximating normal distributions rather than the exponential distributions observed in other models.

Comparing original and comparison datasets, F distributions are nearly identical under the same propagation mode, demonstrating model robustness across different datasets. Interestingly, under the same diffusion mode, F distributions are very similar across the three network structural parameters.

The above figures compare F distributions at threshold 5,000. To further examine distributions across different thresholds, [Figure 10: see original paper] shows the distribution under PSIS model with betweenness centrality propagation for 10 thresholds uniformly selected between 600 and 9,600.

At threshold 600, we observe many users with large influence values, which can be understood as creditors having such low default tolerance that almost every debtor default causes creditor default. When thresholds range from 1,600 to 9,600, user default propagation influence approximately follows a normal distribution, enabling effective user differentiation.

3 Conclusion

This study utilizes actual transaction data from an internet financial platform during May 14, 2015–April 24, 2017 as the original dataset, with corresponding transactions from July 22, 2016–July 22, 2017 as the comparison dataset. By analyzing behavioral characteristics of default transaction propagation, we propose computational models and high-risk user identification algorithms, applying them to both datasets. Comparison with actual data demonstrates good predictive performance, suggesting broad application value for systemic risk prediction and prevention in internet financial platforms.

References

[1] Xiong Yahua, Xiong Yipeng, Li Ting. Research progress on default risk of internet finance P2P customers [J]. Financial Economy, 2015 (12): 70-77.

[2] Wang Shucheng, Tan Zhongming, Chen Yiyun. Review of P2P lending default risk and its contagion assessment [J]. Wuhan Finance, 2017 (6): 40-44.

[3] Wang Qian, Hartmannwendels T. Credit default risk contagion modeling [J]. Journal of Financial Research, 2008 (10): 162-173.

[4] Li Jieting. Modeling and simulation of correlated credit risk contagion based on complex networks [D]. Chengdu: University of Electronic Science and Technology, 2015.

[5] Chen Tingqiang, He Jianmin. Research on credit risk contagion model based on complex networks [J]. Chinese Management Science, 2014, 28 (11): 111-117.

[6] Wang Shenkan. Research on risk management of P2P lending platforms facing borrower default [D]. Beijing: University of International Business and Economics, 2016.

[7] Ding Lan, Luo Pinliang. Research on P2P lending default risk early warning based on Stacking ensemble strategy [J]. Investment Research, 2017, 36 (4): 41-54.

[8] Tang Jianqin. Research on P2P lending borrower default risk measurement based on decision tree algorithm [D]. Changsha: Hunan Normal University, 2016.

[9] Zhou Jie, Liu Zonghua. Epidemic spreading in complex networks [J]. Frontiers of Physics in China, 2008, 3 (3).

[10] Dorogovtsev S N, Mendes J F F. Evolution of networks [J]. Adv. Phys. 2002, 51: 1079–1187.

[11] Newman M E J. Spread of epidemic disease on networks [J]. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics), 2002, 66.

[12] Zhang Zike, Liu Chuang, Zhan Xiuxiu, et al. Dynamics of information diffusion and its applications on complex networks [J]. Physics Reports, 2016, 651.

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