Finance Research Letters 63 (2024) 105308

Contents lists available at ScienceDirect

Finance Research Letters

journal homepage: www.elsevier.com/locate/frl

Network centrality and credit risk: A comprehensive analysis of peer-to-peer lending dynamics

Yiting Liu [a,b], Lennart John Baals [a,b,*], Jörg Osterrieder [a,b], Branka Hadji-Misheva [a]

[a] *Bern Business School, Bern University of Applied Science, Brückenstrasse 73, 3005 Bern, Switzerland*

[b] *Department of High-Tech Business and Entrepreneurship, Faculty of Industrial Engineering and Business Information Systems, University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands*

ARTICLE INFO

Keywords:
Peer-to-Peer lending
Credit-default prediction
Machine Learning
Network centrality

ABSTRACT

This letter analyzes credit risk assessment in the Peer-to-Peer (P2P) lending domain by leveraging a comprehensive dataset from Bondora, a leading European P2P platform. Through combining traditional credit features with network topological features, namely the degree centrality, we showcase the crucial role of a borrower's position and connectivity within the P2P network in determining loan default probabilities. Our findings are bolstered by robustness checks using shuffled centrality features, which further underscore the significance of integrating both financial and network attributes in credit risk evaluation. Our results shed new light on credit risk determinants in P2P lending and benefit investors in capturing inherent information from P2P loan networks.

1. Introduction

Credit risk assessment is a cornerstone of the financial sector, significantly influencing the stability of financial institutions and the broader economy. Yet more complex loan networks, characterized by multifaceted borrowers with varying credit backgrounds, financial behavior and social influences, have exposed limitations to traditional credit assessment methods (Zhou et al., 2019). Conventional metrics and models for credit risk assessment, such as credit scores and statistical methods, have been widely used and validated (Galindo and Tamayo, 2000) but these metrics often assess the risk of individual loans in isolation. Additionally, default rates in conventional Peer-to-Peer (P2P) lending do vary substantially across the different applications ranging from rates of 4% in corporate credit (Galema, 2020) and rates of 4.6% in consumer lending (Emekter et al., 2015) to rates of more then 10% in personal lending (Lyócsa et al., 2022; Dömötör et al., 2023). Specifically, in personal P2P lending, given the information asymmetry between borrowers and lenders (Emekter et al., 2015) and the fact that most loans remain unsecured, the lenders bear considerable credit risk (Li et al., 2022).

Notwithstanding the merit of prior contributions in this area, we thus strengthen a critical aspect in the analysis of loan networks, specifically within the context of personal P2P lending, that links to the concept of degree similarity (Freeman, 2002) in defaulted loans and non-defaulted loans. As a case in point, for a loan i in a loan portfolio $Q = \{1, 2, \ldots, i, \ldots, I\}$, the risk associated with this specific loan, denoted as $R(i)$, is substantially tied to common credit characteristics ($\mathbf{x_i}$), potential similarity with other loans within the portfolio ($\mathbf{s_i}$) and an idiosyncratic factor (ϵ). This loan similarity $\mathbf{s_i}$ can be a vector measuring the status of the loan in the whole portfolio Q. The overall risk of i can be expressed as a function $R(i) = f(\mathbf{x_i}, \mathbf{s_i}, \epsilon)$, where f encapsulates a modeling process

[*] Corresponding author at: Department of High-Tech Business and Entrepreneurship, Faculty of Industrial Engineering and Business Information Systems, University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands.

E-mail addresses: yiting.liu@utwente.nl (Y. Liu), l.j.baals@utwente.nl (L.J. Baals), joerg.osterrieder@utwente.nl (J. Osterrieder), branka.hadjimisheva@bfh.ch (B. Hadji-Misheva).

https://doi.org/10.1016/j.frl.2024.105308
Received 9 January 2024; Received in revised form 11 March 2024; Accepted 25 March 2024
Available online 26 March 2024

used to account for observed and latent factors. Most credit scoring models applied by P2P lending platforms do substantially address individual credit characteristics but fail to account for latent loan similarity factors that are inherent in a network modeling approach. Hence, we argue in this study that by retrieving hidden information located in the network positioning of an individual loan respectively to other loans, we can further improve the classification performance of credit scoring models. Our work proposes a new modeling approach to retrieve information on the distance and degree of defaulted loans and subsequently leveraging this information in the loan default classification process via a step-wise logistic regression model. In doing so we provide new knowledge to the field of credit scoring in P2P lending markets.

Research on loan networks emphasizes their vital role for credit risk assessment and financial stability (Kanno, 2022). Most networks are commonly modeled as directed graphs, where nodes symbolize financial institutions or individual borrowers and edges denote borrowing relationships (Rogers and Veraart, 2013). Seminal works, as in Elsinger et al. (2006) and Rogers and Veraart (2013), have explored systemic risks and crisis management within these networks, thereby highlighting that correlations in banks' asset portfolios are essential risk drivers, and that solvent banks can effectively mitigate cascades of failures. Conversely, the growth of P2P lending platforms has broadened the scope of study within the literature, with Machine Learning (ML) techniques like LightGBM and XGBoost emerging as popular tools for credit risk assessment in this space (Ma et al., 2018).

Contemporary studies focus on model explainability, fairness, and bias mitigation (Tran et al., 2022; Liley et al., 2021), with feature selection and diverse data sources gaining prominence (Trivedi, 2020; Zhou et al., 2021). Notably, while P2P lending studies proliferate (Wang et al., 2022; Zhou et al., 2019), only a subset explores network effects for predicting defaults.

Studies by Giudici et al. (2019) and Giudici et al. (2020) investigate the importance of network topological features in the credit scoring process by utilizing similarity networks in the context of corporate level loan data for a P2P lending network of small- and medium sized enterprises (SMEs). The authors showcase that the inclusion of the network centrality features, namely PageRank (Brin and Page, 1998) and degree centrality imposes superior prediction accuracy to their credit scoring model. Conversely, Chen et al. (2022) are among the first to utilize network centrality measures to assess the influence of network effects on lending and borrowing strategies in the personal P2P lending market, thereby highlighting the importance of lender positioning within the network.

Therefore, we target two core challenges: motivated by the assumption that the credit status of a borrower is portrayed by his or her historical behavior (Liu et al., 2022), we identify the shortcomings of current methods in gauging 'proximity' to loan defaults, and devising a versatile method to effectively capture relational dynamics of such loans (Doumpos et al., 2019) in the personal P2P lending market. Despite studies probing individual risk factors, comprehensive models that untangle the complexities of financial networks remain sparse (Trivedi, 2020; Bhuvaneswari et al., 2014).

Our study aims to fill this gap by introducing a novel methodology that leverages network-induced loan similarity within a P2P loan pool. For this analysis, we utilize a sub-dataset from Bondora, a European P2P lending platform, which has been active since 2009 in Estonia, Finland, Spain, and Slovakia. The platform boasts funds from 225,837 individual lenders and has disbursed €867.5 Mio. in loans.[1] Bondora currently assesses all borrowers via a machine-based scoring process that primarily factors in observable credit features related to credit histories as well as personal- and financial backgrounds.[2] Using the degree centrality and a refined logistic regression model, our method promises not only enhanced risk predictions but also improved model transparency, underscoring the importance of a loan's relative position within a network structure.

2. Research design and data

2.1. Data

The raw dataset was downloaded on April 22nd, 2022, as a part of Bandora's daily updated public report.[3] Loan starting dates span from June 16th, 2009, to April 21st, 2022. The original dataset covers 231,039 individual borrowers characterized through 112 categorical and continuous variables. Among these loans, 79,424 have been recorded with delayed interest payments according to the platform, while 151,615 loans have no recorded delay on interest payments before the download date of the data. Specifically, the dataset details borrower demographics, financial attributes, and past credit market interactions.

2.2. Data cleaning

The data preparation phase is a critical step in transforming the raw dataset into a format that is clean, accurate, and usable for credit risk modeling. This section outlines the conversion of the previously described raw data into cleaned data that is suitable as model input. The process starts with the removal of columns identified as uninformative, such as *DateOfBirth* and *UserName*. The dataset is then reduced to 61 variables. Key date variables such as *ListedOnUTC* and *LoanDate* are converted from string to date formats. Subsequently, we transform categorical variables into dummy variables, and continuous variables are transformed using logarithmic operations to normalize their distributions. A binary *default* indicator is created based on the *DefaultDate* attribute, assigning a value of 1 to loans with a non-null *DefaultDate* and a value of 0 to those without. The default flag follows the platform's

[1] https://www.bondora.com/en

[2] https://help.bondora.com/hc/en-us/articles/15805797702417-How-reliable-is-the-credit-model

[3] https://www.bondora.com/en/public-reports

definition of classifying a loan as defaulted if interest payments are overdue beyond 60 days.[4] The dataset's row count is directly affected through the exclusion of loans based on specific criteria, such as certain *VerificationType* values, *EmploymentStatus* = 0, *Education* = -1, *HomeOwnershipType* = 0, loans with a total income of 0, and loans with negative free cash flow. Additionally, all loans currently marked as *Current* are removed. The dataset is further refined by calculating key indicators, including categorized loan durations, financial ratios, and the Modified Internal Rate of Return (MIRR). This preparation stage ensures the dataset is ready for subsequent analysis and modeling steps. After this initial stage of data cleaning, the dataset has 49,207 rows and 191 columns including *default*.

Afterwards, the dataset undergoes further cleaning to enhance its suitability for the analysis. Rows containing missing values are eliminated to ensure dataset completeness. The dataset is then filtered to retain only rows where *lang.1* equals 1, and subsequently, all columns starting with *lang* are removed to focus on only Estonian loans. This is done due to other 'foreign' loans being considerably riskier reflected in a higher default rate (Dömötör et al., 2023), which would add additional heterogeneity to our modeling process. Date columns, specifically *date.start* and *date.end*, are excluded because the information regarding the length of the loan is already captured by the variable *duration*. In our study, we do not consider the specific start date of the loan to have an impact on the likelihood of default. Forward-looking biased variables, such as *return*, *RR1*, *RR2.Mean*, *RR2.Median*, *RR2.WMean*, *NPRP*, *NPRA*, *FVCI*, *FVCI.Mean*, *FVCI.Median*, and *FVCI.WMean*, which could introduce bias into predictive models by providing information not available at the prediction time, are identified and removed. Duplicate income variables are streamlined, with only those ending in '.no' retained to simplify the dataset. A selection of dummy variables deemed non-essential for analysis, including *AA*, *educ.6*, *em.dur.5p*, *use.m*, *ver.2*, *Mining*, and *Utilities*, are dropped. A correlation matrix is generated for the remaining variables, and any with a correlation coefficient greater than 0.95 are identified and eliminated to address multicollinearity issues, further refining the dataset for subsequent analysis. The dataset, with 32,469 rows and 155 columns, is then suitable for the modeling.

2.3. Data sampling and dataset segmentation

After the data cleaning process mentioned in Section 2.2, there are 20,241 non-defaulted loans and 12,228 defaulted loans in the sample. The non-defaulted/defaulted ratio is 1.66. We randomly select 10,000 negative samples (non-defaulted loans) and 6,000 positive samples (defaulted loans). This selection roughly maintains the original dataset's ratio of positive to negative samples ($\frac{10,000}{6,000} \approx 1.67$) while extracting a sample size convenient for the model training.

We further segment the dataset into a training set, a validation set, and a testing set with sample size ratios of 0.70 : 0.05 : 0.25, respectively. Each dataset maintains the same ratio of non-defaulted to defaulted samples as found in the original dataset. Here, the validation set is not necessary for the logistic model, which is a supervised model used to predict the loan's default status. However, we retain this dataset for potential use with other supervised models that may require hyperparameter tuning.

2.4. Data modeling

To predict the default status of the loan, we propose our degree centrality model. This model contains two steps. The first step constructs two degree centrality features by building a graph on the dataset. The second step applies a supervised model to predict the default status.

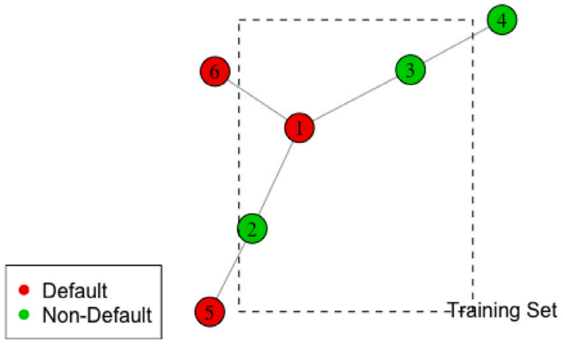
2.4.1. Graph building and degree centrality features constructing

In this section, we introduce the method to build a graph on the dataset and extract degree centrality features. We would stress that this process needs values of all predictors ($\mathbf{x_i}$) in all datasets, including the testing set, while the default status of all loans (y_i) in the testing set are not used to train the model. By this assumption, we reasonably mirror reality. When a borrower applies for a loan, both borrower information and loan details are known at the time of the application and can therefore be used in the prediction. Information unknown *ex ante*, is the potential outcome (defaulted or non-defaulted) of the incoming loan, which the model will predict.

An undirected weighted fully connected graph is employed to measure the similarity of loans. In this graph, each node represents a specific borrower or loan contract. In the Bandora dataset, each observation encapsulates information for one loan contract and its corresponding borrower, making it applicable to represent one loan contract with one node in the graph. Node i is denoted using vector $\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip}, \ldots, x_{iP})'$, where x_{ip} signifies the pth feature among the total P features of this loan. In the fully connected graph, an edge is present between any pair of nodes. Weights are allocated to each edge by computing the Gower's distance (Gower, 1971) between two nodes. The Gower's distance has the advantage of accepting mixed inputs of continuous and categorical variables and measures the similarity between two nodes. The smaller the distance is, the more similar the two nodes are, and vice versa. For each pair of borrowers i and j, the Gower's distance d_{ij} is calculated as:

$$d_{ij} = w_{ij} = \sum_{p=1}^{P} \frac{1}{P} \times \frac{d_{ij}^{p}}{\max(x_{\cdot p}) - \min(x_{\cdot p})}, \text{where } d_{ij}^{p} = \begin{cases} |x_{ip} - x_{jp}| & \text{if } x_p \text{ is a continuous variable,} \\ 1 & \text{if } x_p \text{ is a categorical variable and } x_{ip} \neq x_{jp}, \\ 0 & \text{if } x_p \text{ is a categorical variable and } x_{ip} = x_{jp}. \end{cases}$$

[4] https://www.bondora.com/blog/p2p-finance-association-defaults/. In turn, the P2P Finance Association reports a loan default definition of 120 days as industry standard (https://p2pfa.info).



node i	$\deg^0(i)$	$\deg^1(i)$
1	2	0
2	1	0
3	1	0
4	0	1
5	0	1

Fig. 1. A Simplified Example on the Calculation of $\deg^0(i)$ and $\deg^1(i)$.

Subsequently, this fully connected graph is reduced to its Minimum Spanning Tree (MST) (Prim, 1957). In the MST, we count how many defaulted nodes and non-defaulted nodes in the training set a specific node is linked to. The numbers of defaulted nodes and non-defaulted nodes that one specific node is linked to are denoted as $\deg^1(i)$ and $\deg^0(i)$, respectively. Here we include a simplified example to illustrate the calculation (Fig. 1).

In this case, node $i = 1$ is connected to two non-defaulted points, node $i = 2$ and node $i = 3$. Node $i = 1$ is also connected to a defaulted node $i = 6$. However, node $i = 6$ is not in the training set, thus the default status of $i = 6$ cannot be known from an ex ante perspective. The result is $\deg^1(1) = 0$ and $\deg^0(1) = 2$. The calculation is similar for node $i = 2$ and node $i = 3$. As for node $i = 4$, its own default status cannot be known, but it is linked to node $i = 3$ that is included in the training set. Thus, $\deg^0(3) = 1$. The calculation is similar for node $i = 5$. From this example, we can infer that the default status of nodes not included in training set are consequently never involved in the model training, which is consistent with reality. Besides, our computation is defined by the calculation of degree centrality in graph theory (Friedkin, 1991). Hence, we still use “deg” to name the variables we calculate following this approach. A higher value of $\deg^0(i)$ means that the node is closer to non-defaulted loans in the graph, while a higher value of $\deg^1(i)$ means the node is closer to defaulted loans.

2.4.2. Model development and evaluation

Upon the completion of feature engineering, a step-wise logistic regression model is developed to predict the binary outcome variable y_i of a loan. The models are trained using the training dataset and evaluated on the testing dataset, with a particular focus on assessing the impact of the degree centrality features on predictive performance.

The logistic regression model is expressed mathematically as:

$$logit(P(y_i = 1)) = \beta_0 + \boldsymbol{\beta}'\mathbf{x_i} + \beta^1 \deg^1(i) + \beta^0 \deg^0(i) + \epsilon, \tag{1}$$

where $P(y_i = 1)$ represents the probability of default, β_0 is the intercept, $\boldsymbol{\beta}$ is the coefficient vector for loans' own credit characteristics, β^1 and β^0 are coefficients for centrality measures, and ϵ is an error term respectively. In this form we are able to capture a linear relationship between the values of the variables and the log-odds (Dömötör et al., 2023), which is beneficial to our study approach. The significance of β^1 and β^0 indicates that the two centrality measures are important in prediction.

To further confirm the contribution of these features on prediction accuracy, we run the model without centrality measures as a baseline model in parallel. That is:

$$logit(P(y_i = 1)) = \beta_0 + \boldsymbol{\beta}'\mathbf{x_i} + \epsilon, \tag{2}$$

where parameters and variables have the same definitions as in Model (1).

In the initial phase of the model training, we conduct feature selection, utilizing a Random Forest model. The Gini importance measure is employed to select a subset of features, which are then used as predictors in the logistic regression model. The Random Forest selects the most important 20 features among 159 features. Based on the 20 features, a step-wise feature selection procedure is then applied, which iteratively adds and removes predictors to find a model that balances fit and complexity, thereby optimizing the Akaike Information Criterion (AIC) (Akaike, 2011). This two-stage feature selection process, strategically combines the strengths of Random Forest and step-wise algorithm to improve variable selection efficiency. By initially using a Random Forest with 500 trees for its ability to handle complex interactions and identify key variables from a large pool, we narrow down the variable set to the most impactful ones. These selected variables are then analyzed using logistic regression, capitalizing on its interpretability and ability to estimate linear relationships. This approach not only streamlines the selection process, reducing the risk of overfitting and enhancing model generalizability, but also ensures that the variables included in the final model are those with substantial predictive power, leading to a more refined and effective logistic regression model.

Table 1
Model output for the logistic regression without degree centrality feature. The table presents the coefficient estimates, standard errors, z values, and p-values for each predictor variable in the logistic regression model without the inclusion of degree centrality features. The significance of the predictors can be evaluated based on the p-values (Pr(> |z|)). Listed variables are defined in Appendix.

Variable	Estimate	Std. error	z value	Pr(> |z|)
duration.60	0.907	0.074	12.235	<0.001
duration.36	0.880	0.072	12.259	<0.001
log.amount	0.247	0.042	5.947	<0.001
NoOfPreviousLoansBeforeLoan	0.141	0.035	4.027	<0.001
DebtToIncome	0.113	0.041	2.764	0.006
Interest	0.037	0.023	1.635	0.102
inc.princ.empl.l	−0.111	0.076	−1.459	0.145
AmountOfPreviousLoansBeforeLoan	−0.117	0.037	−3.193	0.001
MonthlyPayment	−0.154	0.039	−3.897	<0.001
homeless	−0.168	0.043	−3.918	<0.001
inc.total	−0.177	0.024	−7.348	<0.001
ver.4	−0.391	0.046	−8.421	<0.001
FreeCash.l	−0.397	0.061	−6.479	<0.001
educ.4	−0.498	0.052	−9.628	<0.001
(Intercept)	−0.512	0.079	−6.484	<0.001
time	−0.653	0.036	−17.909	<0.001
duration.12	−0.763	0.161	−4.728	<0.001
educ.5	−1.080	0.064	−16.837	<0.001

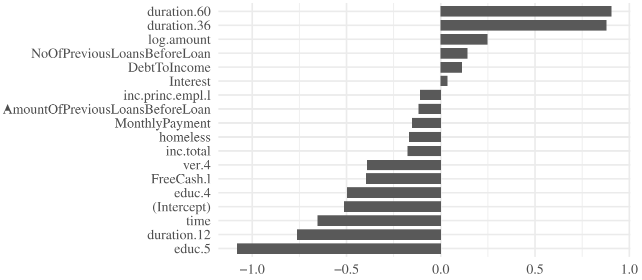


Fig. 2. Feature Importance for the Logistic Regression Model without Degree Centrality on the Training Dataset.

3. Results

3.1. Basic model calibration

The initial benchmark model in form of the step-wise logistic regression, which excludes degree centrality features, utilizes various credit features initially selected from a Random Forest model. This model utilizes the set of credit features without centrality measures. As shown in Table 1, the step-wise selection process refines the logistic regression, retaining significant economic predictors such as time (*time*), total borrower income (*inc.total*), loan payments (*MonthlyPayment*), and interest payments (*Interest*), among others, with an AIC of 13174.06. These predictors, primarily financial attributes, confirm the baseline understanding of the credit factors, influencing loan default status in the P2P lending market (Lyócsa et al., 2022; Giudici et al., 2020).

The feature importance depicted in Fig. 2, for the logistic regression model, excluding degree centrality features, further confirms the insights into credit features that significantly influence the likelihood of loan default in the P2P lending market.

In line with previous work, we find a common subset of credit features, namely extensive loan duration (*duration.601*, *duration.361*) (Ma et al., 2018), loan amount borrowed (*log.amount*) and interest payments (*Interest*) (Babaei and Bamdad, 2020; Wu and Zhang, 2021), the previous number of loans borrowed (*NoOfPreviousLoansBeforeLoan*), and the borrower's leverage ratio (*DebtToIncome*) (Giudici et al., 2020) to exhibit a positive association with default likelihood. Similarly, we find that an increase in these variables is associated with a heightened probability of loan default. Our findings further coincide with other scholarly work that finds loan-induced information to be among the most influential features to impact credit default (Lee et al., 2021; Lyócsa et al., 2022). For instance, an elongation in loan duration (*duration.601* and *duration.361*) and an augmentation in the log-transformed loan amount (*log.amount*) are correlated with an elevated risk of default, underscoring the risk associated with longer-term and larger-sized loans. Similarly, different levels of interest rates, influencing loan repayment, borrowing history, and the leverage ratio of the borrower do positively influence loan default probabilities, which is in line with our expectation.

Conversely, individual borrower-related features like the type of income (*inc.princ.empl.l*), education type (*educ.41*; *educ.51*), liquidity bolster (*FreeCash.l*), time elapsed (*time*), loan application verification (*ver.41*), and no home ownership (*homeless1*) are negatively associated with default likelihood. We subsume a potential reason for this variable importance is due to the fact that

Table 2
Model output for the logistic regression with degree centrality feature. The table presents the coefficient estimates, standard errors, z values, and p-values for each predictor variable in the logistic regression model with the inclusion of degree centrality features. The significance of the predictors can be evaluated based on the p-values (Pr(> |z|)). Listed variables are defined in Appendix.

Variable	Estimate	Std. error	z value	Pr(> |z|)
duration.60	0.848	0.075	11.320	<0.001
duration.36	0.830	0.072	11.449	<0.001
log.amount	0.252	0.042	6.045	<0.001
degree_pos	0.165	0.021	7.710	<0.001
NoOfPreviousLoansBeforeLoan	0.128	0.035	3.601	0.000
DebtToIncome	0.070	0.031	2.309	0.021
Interest	0.034	0.023	1.532	0.126
AmountOfPreviousLoansBeforeLoan	−0.112	0.037	−3.026	0.002
degree_neg	−0.136	0.022	−6.166	<0.001
MonthlyPayment	−0.154	0.039	−3.908	<0.001
inc.total	−0.169	0.025	−6.981	<0.001
ver.4	−0.339	0.047	−7.289	<0.001
FreeCash.l	−0.446	0.037	−12.024	<0.001
educ.4	−0.477	0.052	−9.195	<0.001
(Intercept)	−0.601	0.078	−7.695	<0.001
time	−0.618	0.036	−16.919	<0.001
duration.12	−0.734	0.162	−4.544	<0.001
educ.5	−1.004	0.065	−15.512	<0.001

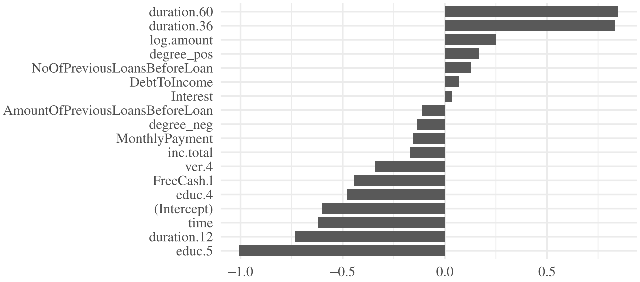


Fig. 3. Feature Importance for the Logistic Regression Model with Degree Centrality on the Training Dataset.

personal P2P lending is majorly conducted by younger aged individuals with less income and fewer guarantees (Jiang et al., 2018). With increasing wealth through career progression for most young borrowers the magnitude of these variables becomes more prominent, thereby signaling economic strengths to lenders. Accordingly, an increase in the amount of these borrower-related 'soft' features is found to significantly reduce the probability of loan defaults (Jiang et al., 2018; Lee et al., 2021; Trivedi, 2020). For instance, an increase in the total income (*inc.total*) and cash liquidity (*FreeCash.l*) of a borrower is typically associated with a decreased propensity for loan default, reflecting financial stability.

3.2. Incorporating degree centrality features into the model

The model incorporating degree centrality features is capable to additionally capture latent factors embedded in the designed similarity network that we obtained from the borrowers' feature distances. From Table 2 both degree centrality features computed on defaulted (*degree_pos*) and non-defaulted (*degree_neg*) loans are retained as significant predictors in the model. We infer from this that borrowers with higher positive degree centrality (higher similarity to defaulted loans) and lower negative degree centrality (lower similarity to non-defaulted loans) are more prone to default. More specifically a higher positive degree centrality of an individual borrower signals closer location to defaulted borrowers, as their credit and demographic features rank similarly to defaulted peers. Conversely, a higher negative degree centrality indicates closer positioning to non-defaulted borrowers, revealing similar characteristics of solvent borrowers. Hence, the topological information hidden in the network highlights the borrower's connectivity, defined by his or her credit, personal, and financial characteristics, in contrast to other borrowers. With an AIC of 13 149.57, the model achieves a comparable goodness of fit to the model excluding degree centrality features.

Notably, as can be seen in Fig. 3, the magnitude of $\deg^1(i)$ (*degree_pos*) or $\deg^0(i)$ (*degree_neg*), indicates the positive or negative impact that a borrower's connection with defaulted and non-defaulted peers has.

Our graph-enhanced model depicts the same subset of conventional credit features (loan duration (*duration.601*, *duration.361*), loan amount borrowed (*log.amount*), previous number of loans borrowed (*NoOfPreviousLoansBeforeLoan*), borrower's leverage ratio (*DebtToIncome*), and interest payment (*Interest*) as positive contributors to default risk in contrast to the conventional model. Conversely, individual borrower-related features such as the Amount of previous loans borrowed (*AmountOfPreviousLoansBeforeLoan*), and the total income of a debtor (*inc.total*) appear to be negatively correlated with default likelihood.

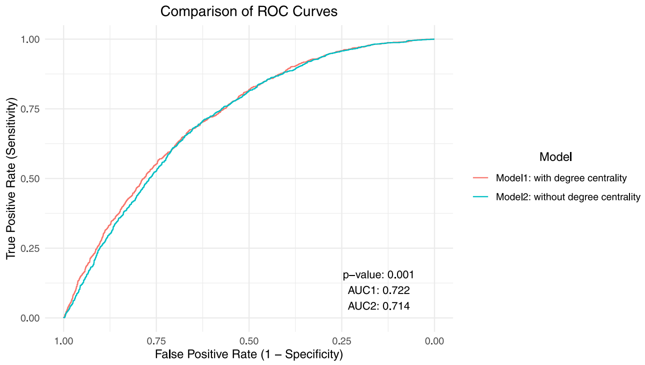


Fig. 4. Comparison of ROC Curves for Logistic Regression Models with and without Degree Centrality.

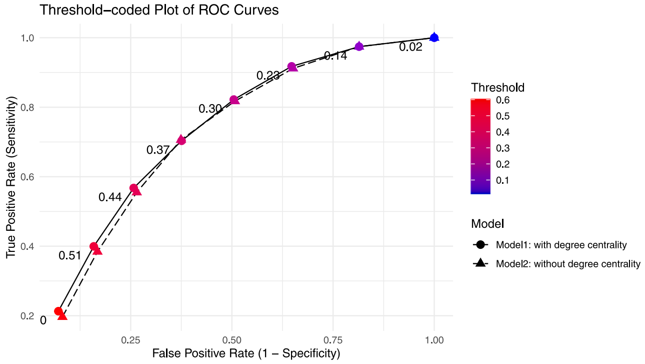


Fig. 5. Comparison of Threshold-Coded ROC Curves for Logistic Regression Models with and without Degree Centrality.

The key distinction between both models lies in the impact of the computed degree centrality features. The degree centrality on defaulted loans (*degree_pos*) exhibits a positive influence on loan default likelihood. Contrarily, the degree centrality feature on non-defaulted loans (*degree_neg*) negatively affects the default likelihood of the borrower. Hence, our model underscores that by discriminating between defaulted and non-defaulted borrowers, and accounting for their interconnectedness through a similarity network, further information is utilzed for enhanced borrower default prediction.

3.3. Comparative analysis of models

While both logistic regression models demonstrate similar goodness of fit, the inclusion of degree centrality features improves the credit scoring as can be seen in Fig. 4. The significant coefficients of the degree centrality features (*degree_pos*; *degree_neg*) underscore their utility in enhancing the predictive capabilities of the model.

As can be seen in Fig. 5, the model with the degree centrality features is able to capture more true positives (defaults) relative to a given level of false positives (non-default) in contrast to the model with only conventional credit features. For higher thresholds implying stronger certainty of credit default, the network-enhanced model can yield additional assurance in identifying default cases. Though this effect diminishes at higher levels of uncertainty, the model does not perform worse, indicating a reasonable benefit of the graph-based architecture. With this effect we observe a similar improvement in loan default classification performance as in Lee et al. (2021), where graph-level prediction of the entire graph feature from node (borrower) and edge (similarity) information is utilized. Thus, our integration of degree centrality as graph features in the credit risk model emphasises the importance of considering topological attributes in addition to established financial variables in risk assessment. These findings coincide with similar scholarly works in the field of credit default prediction on personal- and corporate loans (Giudici et al., 2020; Yıldırım et al., 2021; Ahelegbey et al., 2019). Based on the universal integration of our node-level classification approach, P2P lending platforms, aside from Bondora, can use this method seamlessly as an add-on tool to whatever risk assessment framework they utilize without fundamentally re-training the underlying model. Specifically, in personal P2P lending, where high default rates are observed, our study outcome will allow platforms to further diminish risks associated with the operational efficacy within high-risk lending markets. The significant predictive power of the degree centrality variables suggests that these network features can effectively supplement traditional credit features, by providing a more nuanced and robust basis for evaluating borrower risk in P2P lending platforms.

Table 3
Model output for the logistic regression with shuffled degree centrality feature. The table presents the coefficient estimates, standard errors, z values, and p-values for each predictor variable in the logistic regression model with shuffled degree centrality feature. The significance of the predictors can be evaluated based on the p-values (Pr(> |z|)). Listed variables are defined in Appendix.

Variable	Estimate	Std. error	z value	Pr(> |z|)
duration.60	0.901	0.074	12.177	<0.001
duration.36	0.874	0.072	12.189	<0.001
log.amount	0.247	0.042	5.950	<0.001
NoOfPreviousLoansBeforeLoan	0.139	0.035	3.958	<0.001
DebtToIncome	0.114	0.041	2.792	0.005
Interest	0.037	0.023	1.652	0.099
inc.princ.empl.l	−0.110	0.076	−1.446	0.148
AmountOfPreviousLoansBeforeLoan	−0.119	0.037	−3.234	<0.001
MonthlyPayment	−0.155	0.039	−3.940	<0.001
inc.total	−0.176	0.024	−7.312	<0.001
ver.4	−0.382	0.046	−8.259	<0.001
FreeCash.l	−0.398	0.061	−6.481	<0.001
educ.4	−0.503	0.052	−9.734	<0.001
(Intercept)	−0.577	0.077	−7.474	<0.001
time	−0.649	0.036	−17.844	<0.001
duration.12	−0.766	0.161	−4.749	<0.001
educ.5	−1.089	0.064	−16.987	<0.001

3.4. *Robustness checks*

To ensure the robustness of our model estimates, we incorporate two shuffled centrality features into the data sample and model training process. These features are derived from the degree centrality measures for both the positive defaulted and negative non-defaulted groups, with their structure being randomly rearranged to nullify any dependencies. Thus, the inclusion of the shuffled centrality features should not offer more predictive power for classifying loan defaults than a randomly drawn i.i.d sequence (Dimpfl and Peter, 2018). Subsequently, we train the step-wise logistic regression model on data sets that solely included the conventional credit features, paired with the randomized degree centrality features, to evaluate the model performance under the respective feature selection. Under our *apriori* assumption, none of the models, when trained with the inclusion of the randomized centrality features, should reveal any significant feature importance in the default classification process. In addition, we also separately repeate the estimation step of the graph-based modeling approach by replacing the step-wise logistic regression model with a non-parametric model, here a Random Forest with 150 trees and two variables randomly selected, to ensure the validity of the observed feature importance. The results confirm the chosen features, initially categorized by the step-wise logistic regression model, to hold similar or equal importance in the Random Forest (see Table 3). [5]

4. Conclusion

In this letter, we model credit risk within personal Peer-to-Peer (P2P) lending, using a dataset from the European platform Bondora. By integrating conventional credit features with network centrality measures, specifically degree centrality, we highlight the importance of a borrower's position and connectivity within the P2P loan pool for predicting default likelihood. Our findings emphasis the importance of considering graph-related features in credit risk modeling, where the borrower base is heterogeneous and topological information is available. Robustness checks using shuffled degree centrality features and an alternative non-parametric modeling approach, validate the significance of the predictive improvement in loan default classification in our original analysis. Personal P2P lending remains characterized by default risk primarily borne by the investor, respectively the lender, and not by the lending platform itself. Thus, our findings emphasize the value of combining financial and network attributes in credit risk assessment, thereby suggesting a more comprehensive approach to investors in understanding borrower risk and enhancing accurate credit risk scoring in P2P lending platforms.

CRediT authorship contribution statement

Yiting Liu: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lennart John Baals:** Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jörg Osterrieder:** Validation, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Branka Hadji-Misheva:** Validation, Supervision, Resources, Project administration, Funding acquisition, Formal analysis, Conceptualization.

[5] Detailed results of the analysis are available upon request from the authors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors want to thank Štefan Lyócsa, Tomáš Plíhal (Masaryk University), Ali Hirsa (Columbia University), and COST Action CA19130 Management Committee and Working Group members for their invaluable help and comments on the research design of this study.

Funding

This work has been supported by several institutions: COST Action CA19130 and COST Action CA21163, under the auspices of the European Cooperation in Science and Technology (COST). Swiss National Science Foundation under multiple projects: Mathematics and Fintech (IZCNZ0-174853), Anomaly and Fraud Detection in Blockchain Networks (IZSEZ0-211195), Narrative Digital Finance: a tale of structural breaks, bubbles & market narratives (IZCOZ0-213370).

Notably, we are grateful for financial support from the Swiss National Science Foundation under the project Network-based credit risk models in P2P lending markets (100019E-205487).

We gratefully acknowledge the support of the Marie Skłodowska-Curie Actions under the European Union's Horizon Europe research and innovation program for the Industrial Doctoral Network on Digital Finance, acronym: DIGITAL, Project No. 101119635.

In addition, our research has benefited from funding from the European Union's Horizon 2020 research and innovation program under the grant agreement No 825215 (Topic: ICT-35-2018, Type of action: CSA).

Support from the Leading House Asia under the project Graph-Theoretic Analysis for Consumer Credit Risk Assessment in Personal Lending is greatly appreciated.

Furthermore, we acknowledge the ongoing cooperation between the ING Group and the University of Twente.

Appendix. Loan application data features and descriptions

Features	Description
date.start	Date when the loan was issued
date.end	Loan maturity date according to the latest loan schedule
default	1 - loan defaulted, 0 - otherwise
return	Nominal annual return $= [(\text{Future value of all cash-inflows} + \text{loan amount})/\text{loan amount}]^{1/(\text{actual loan duration in days}/365)} - 1$
RR1	Modified Internal Rate of Return - 0 re-investment rate
RR2	Modified Internal Rate of Return - re-investment rate given by return on loans ending 365 days prior to the start of the loan (mean, median, weighted)
NPRP	Nominal profit in % = cash inflows/loan amount - 1
NPRA	Nominal net cash = sum of all cash inflows - loan amount
FVCI	Future value of cash inflows - loan amount, re-invested at the return (mean, median, weighted)
new	1 - it is a new customer
Age	The age of the loan applicant
Gender	1 - Woman, 0 - Male or couple of undefined
Interest	Maximum interest accepted in the loan application
MonthlyPayment	Estimated amount the borrower has to pay every month
No. Prev. Loans	Number of previous loans
Amt. Prev. Loans Bef. Loan	Value of previous loans
time	Time index in days = Current date of the loan application - Earliest date of a loan application in the dataset
time2	Square of the time index
time3	Cube of the time index
Hour	Application hour (ranging from 0 to 22)
weekday	Day of the week (1 for Friday, 2 for Monday, 3 for Saturday, 4 for Sunday, 5 for Thursday, 6 for Tuesday)
ver	Method used to verify loan application data (2 for income unverified and cross-referenced by phone, 3 for income verified, 4 for income and expenses verified)

Features	Description
lang	Language (1 for Estonian, 2 for English, 3 for Russian, 4 for Finnish, 6 for Spanish)
log.amount	Natural log of the loan amount
duration	Duration of the loan in months (options include 6, 9, 12, 18, 24, 36, 48, 60 months)
use	Loan use - consolidation, real estate, home improvement, business, education, travel, vehicle, other, health, not specified
educ	Loan applicant's education - basic education, vocational education, secondary education, higher education, not specified
marital	Loan applicant's marital status - married, cohabitant, single, divorced, widow
depen	Loan applicant's number of children or other dependents - 0, 1, 2, 3, 4
employ	Loan applicant's employment status - partially employed, fully employed, self-employed, entrepreneur, retiree
em.dur	Loan applicant's employment duration - more than 5 years, other, retiree, trial period, less than 1 year, less than 2 years, less than 3 years, less than 4 years, less than 5 years
exper	Loan applicant's experience - less than 2 years, less than 5 years, less than 10 years, less than 15 years, less than 25 years, more than 25 years
Other	Loan applicant's occupation area
Mining	Loan applicant's occupation area
Processing	Loan applicant's occupation area
Energy	Loan applicant's occupation area
Utilities	Loan applicant's occupation area
Construction	Loan applicant's occupation area
Retail.wholesale	Loan applicant's occupation area
Transport.warehousing	Loan applicant's occupation area
Hospitality.catering	Loan applicant's occupation area
Info.telecom	Loan applicant's occupation area
Finance.insurance	Loan applicant's occupation area
Real.estate	Loan applicant's occupation area
Research	Loan applicant's occupation area
Administrative	Loan applicant's occupation area
Civil.service.military	Loan applicant's occupation area
Education	Loan applicant's occupation area
Healthcare.social.help	Loan applicant's occupation area
Art.entertainment	Loan applicant's occupation area
Agriculture.for.fish	Loan applicant's occupation area
homeless	Loan applicant's home ownership type - homeless
owner	Loan applicant's home ownership type - owner
livingw.parents	Loan applicant's home ownership type - living with parents
tenant.pfp	Loan applicant's home ownership type - tenant, pre-furnished property
council.house	Loan applicant's home ownership type - council house
joint.tenant	Loan applicant's home ownership type - tenant
joint.ownership	Loan applicant's home ownership type - joint ownership
mortgage	Loan applicant's home ownership type - mortgage
encumbrance	Loan applicant's home ownership type - owner with encumbrance
inc.princ.empl.no	1 - has income from a principal employer
inc.princ.empl.l	1 - The amount of income from the principal employer [log(x+1)]
inc.pension.no	1 - has income from a pension
inc.fam.all.no	1 - has income from family allowances
inc.soc.wel.no	1 - has income from social welfare
inc.leave.no	1 - has income from leave
inc.child.no	1 - has income from child support
inc.other.no	1 - has income from other sources
inc.total	Total income [log(x+1)]
no.liab	Loan applicant's number of existing liabilities (0, 1, 2, 3, 4, 5, up to 10)
liab.l	Total amount of existing liabilities [log(x+1)]
no.refin	Loan applicant's number of liabilities after refinancing (0, 1, 2, 3, 4)
inc.support	Loan applicant's income from alimony payments [log(x+1)]
FreeCash.d	1 - has free cash
FreeCash.l	Total amount of free cash [log(x+1)]
no.previous.loan	Loan applicant's number of previous loans (0, 1, 2, 3, 4, 5, 6, 7)
previous.loan.l	Total amount of loan applicant's previous loan amounts [log(x+1)]
no.previous.repay	Loan applicant's number of previous early repayments (0, more than 1)
previous.repay.l	Total amount of loan applicant's previous loan repayments [log(x+1)]
A	Bondora rating - A
AA	Bondora rating - AA
B	Bondora rating - B
C	Bondora rating - C

References

Ahelegbey, D.F., Giudici, P., Hadji-Misheva, B., 2019. Latent factor models for credit scoring in P2P systems. Physica A 522, 112–121.

Akaike, H., 2011. Akaike's information criterion. In: International Encyclopedia of Statistical Science. p. 25.

Babaei, G., Bamdad, S., 2020. A multi-objective instance-based decision support system for investment recommendation in peer-to-peer lending. Expert Syst. Appl. 150, 113278.

Bhuvaneswari, U., Paul, P.J.D., Sahu, S., 2014. Financial risk modelling in vehicle credit portfolio. In: 2014 International Conference on Data Mining and Intelligent Computing, ICDMIC 2014.

Brin, S., Page, L., 1998. The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. 30 (1–7), 107–117.

Chen, X., Chong, Z., Giudici, P., Huang, B., 2022. Network centrality effects in peer to peer lending. Physica A 600, 127546.

Dimpfl, T., Peter, F.J., 2018. Analyzing volatility transmission using group transfer entropy. Energy Econ. 75, 368–376.

Dömötör, B., Illés, F., Ölvedi, T., 2023. Peer-to-peer lending: Legal loan sharking or altruistic investment? Analyzing platform investments from a credit risk perspective. J. Int. Financ. Mark. Inst. Money 86, 101801.

Doumpos, M., Lemonakis, C., Niklis, D., Zopounidis, C., 2019. Data analytics for developing and validating credit models. In: EURO Advanced Tutorials on Operational Research, pp. 43–75.

Elsinger, H., Lehar, A., Summer, M., 2006. Risk assessment for banking systems. Manage. Sci. 52 (9), 1301–1314.

Emekter, R., Tu, Y., Jirasakuldech, B., Lu, M., 2015. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. Appl. Econ. 47 (1), 54–70.

Freeman, L.C., 2002. Centrality in social networks: Conceptual clarification. In: Social Network: Critical Concepts in Sociology, vol. 1, Routledge, Londres, pp. 238–263.

Friedkin, N.E., 1991. Theoretical foundations for centrality measures. Am. J. Sociol. 96 (6), 1478–1504.

Galema, R., 2020. Credit rationing in P2P lending to SMEs: Do lender-borrower relationships matter? J. Corporate Finance 65, 101742.

Galindo, J., Tamayo, P., 2000. Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. Comput. Econ. 15 (1–2), 107–143.

Giudici, P., Hadji-Misheva, B., Spelta, A., 2019. Network based scoring models to improve credit risk management in peer to peer lending platforms. Front. Artif. Intell. 2, 3.

Giudici, P., Hadji-Misheva, B., Spelta, A., 2020. Network based credit risk models. Qual. Eng. 32 (2), 199–211.

Gower, J.C., 1971. A general coefficient of similarity and some of its properties. Biometrics. J. Int. Biometric Soc. 857–871.

Jiang, C., Wang, Z., Wang, R., Ding, Y., 2018. Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. Ann. Oper. Res. 266 (1–2), 511–529.

Kanno, M., 2022. Exploring risks in syndicated loan networks: Evidence from real estate investment trusts. Econ. Model. 115, 105953.

Lee, J.W., Lee, W.K., Sohn, S.Y., 2021. Graph convolutional network-based credit default prediction utilizing three types of virtual distances among borrowers. Expert Syst. Appl. 168, 114411.

Li, T., Kou, G., Peng, Y., Yu, P.S., 2022. An integrated cluster detection, optimization, and interpretation approach for financial data. IEEE Trans. Cybern. 52 (12), 13848–13861.

Liley, J., Emerson, S.R., Mateen, B.A., Vallejos, C.A., Aslett, L.J., Vollmer, S.J., 2021. Model updating after interventions paradoxically introduces bias. In: Proceedings of Machine Learning Research, vol. 130, pp. 3916–3924.

Liu, J., Zhang, S., Fan, H., 2022. A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network. Expert Syst. Appl. 195, 116624.

Lyócsa, S., Vašaničová, P., Hadji Misheva, B., Vateha, M.D., 2022. Default or profit scoring credit systems? Evidence from European and US peer-to-peer lending markets. Financ. Innov. 8 (1), 32.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., Niu, X., 2018. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. Electron. Commer. Res. Appl. 31, 24–39.

Prim, R.C., 1957. Shortest connection networks and some generalizations. Bell Syst. Tech. J. 36 (6), 1389–1401.

Rogers, L., Veraart, L., 2013. Failure and rescue in an interbank network. Manage. Sci. 59 (4), 882–898.

Tran, K.L., Le, H.A., Nguyen, T.H., Nguyen, D.T., 2022. Explainable machine learning for financial distress prediction: Evidence from vietnam. Data 7 (11).

Trivedi, S.K., 2020. A study on credit scoring modeling with different feature selection and machine learning approaches. Technol. Soc. 63.

Wang, C., Liu, Q., Li, S., 2022. A two-stage credit risk scoring method with stacked-generalisation ensemble learning in peer-to-peer lending. Int. J. Embed. Syst. 15 (2), 158–166.

Wu, Y., Zhang, T., 2021. Can credit ratings predict defaults in peer-to-peer online lending? Evidence from a Chinese platform. Finance Res. Lett. 40, 101724.

Yıldırım, M., Okay, F.Y., Özdemir, S., 2021. Big data analytics for default prediction using graph theory. Expert Syst. Appl. 176, 114840.

Zhou, J., Li, W., Wang, J., Ding, S., Xia, C., 2019. Default prediction in P2P lending from high-dimensional data based on machine learning. Physica A 534.

Zhou, J., Wang, C., Ren, F., Chen, G., 2021. Inferring multi-stage risk for online consumer credit services: An integrated scheme using data augmentation and model enhancement. Decis. Support Syst. 149.