Leveraging Network Topology for Credit Risk Assessment in P2P Lending

COST FinAI Meets Istanbul

Yiting Liu
Lennart John Baals
Branka Hadji Misheva
Joerg Osterrieder

Table of Contents

1. Introduction to P2P Lending and the Bondora Dataset

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

Introduction to P2P Lending - Overview

Definition of P2P Lending

Peer-to-peer (P2P) lending is a method of debt financing that enables individuals to borrow and lend money without the use of an financial institution as an intermediary.

Online Platform Character

P2P lending occurs through online platforms that pair lenders with potential borrowers.

These platforms used to offer more inclusiveness and better funding efficiency than banks.

Benefits for Lenders

Lenders can earn higher returns compared to traditional savings and investment products.

Lenders can choose which borrowers to invest in.

Benefits for Borrowers

Borrowers can access financing faster and often with less stringent credit checks.

This makes it beneficial for personal loan seeking.

Growing Importance

With the power of technology and the internet, P2P lending has grown in popularity since it started in the early 2000s. Today, P2P lending platforms have facilitated billions of dollars in loans *(Chen, Huang, and Shaban 2022)*.

Advantages of P2P Lending Platforms

An Overview

Peer-to-peer (P2P) lending platforms have reshaped the financial landscape.

Micro Perspective

Accessibility: Simplified loan access for individuals without strigent credit checks.

Potential Returns: Can offers higher returns for investors.

Flexibility: More customizable loan terms based on borrower needs.

Macro Perspective

Economic Growth: Stimulates economy through capital flow.

Financial Inclusion: Serves borrowers in areas that are underserved by banks.

Innovation: Disrupts traditional banking and lending, thus driving innovation.

Disadvantages of P2P Lending Platforms

An Overview

While beneficial, P2P lending platforms also have drawbacks and risks for lenders and borrowers.

Micro Perspective

Credit Risk: Higher default risk due to relaxed credit checks.

Limited Insurance: Most P2P loans are uninsured, increasing risk for lenders.

Liquidity Risk: Difficulty in withdrawing investment before loan matures if no buyback guarantee established.

Macro Perspective

Regulatory Uncertainty: New financial model with sparse regulatory frameworks.

Systemic Risk: With platform growth the systemic risk component increases.

Uneven Market: May exacerbate inequality by favoring certain demographic groups.

Introduction to P2P Lending - Overview Part 2

Challenges in P2P Lending

In this context, P2P lending also imposes challenges to creditors as issued loans are usually unsecured which results in higher default rates, and a lower number of recovery options.

Need for Regulation

Given its nature, there's also a need for regulation to protect both borrowers and lenders from increased default risk.

Thus, governments and regulatory bodies are increasingly focusing on P2P lending platforms.

Need for Efficient Credit Scoring

Credit scoring is an important part of P2P lending.

It is used to assess the risk associated with a given borrower.

Integration of Machine Learning

Machine learning is now being used to enhance credit scoring processes.

The potential to improve loan default predictions is immense but yet sparsely investigated.

Summary

In summary, P2P lending is an innovative and growing field, but one that requires careful risk management. The use of machine learning could be a key to improving these risk assessments.

Traditional vs P2P Lending

Overview

In comparison, Peer-to-Peer (P2P) lending represents a significant shift from traditional lending practices.

Traditional Lending

Centralized: Banks act as intermediaries and guide the lending process.

Risk Assessment: Banks use internal credit scoring models that have rich debtor information at hand.

Interest Rates: Determined by the bank, often opaque.

Regulation: Subject to extensive regulation.

P2P Lending

Decentralized: Platform connects borrowers and lenders directly.

Risk Assessment: Focus on publicly available credit ratings and alternative data sources to apply innovative methods.

Interest Rates: Determined by the P2P lending platform based on credit scoring.

Regulation: Less regulated, but increasing.

Significance

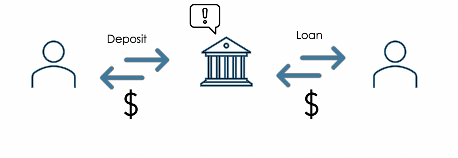
The unique characteristics of P2P lending, such as its decentralized nature and publicly available credit information, open up new opportunities for credit risk modeling based on network analysis.

P2P Lending: Forms and Risk Ownership

Traditional Lending

Risk Ownership: The institution (bank) that provides the credit score also takes on the risk of the loan defaulting.

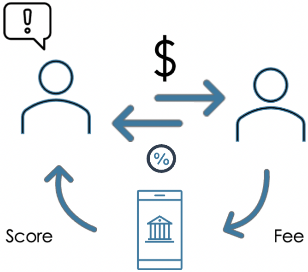
Incentives: Banks have an inherent interest in providing accurate credit scoring due to the direct risk they bear.



P2P Lending

Risk Ownership: The platform provides the credit score, but the credit risk is borne by the investors.

Incentives: The platform's primary interest lies in increasing lending volumes, which may compromise the accuracy of credit scoring.



Bondora: Platform and Dataset Processing

The Platform

Bondora (`https://www.bondora.com/en`) is a peer-to-peer (P2P) lending platform:

- Established in Estonia in 2008; Serving over 226,696 customers;
- Over €902 million invested; More than €113 million paid out in interest
- (all data collected on December 7th, 2023).

Clean the data:

- Select loan records only in Estonian; Delete records with NA values.
- Drop variables:
  - Irrelevant variables: `date.start`, `date.end`
  - Variables unknown before a loan ends: `return`, `RR1`, `RR2.Mean`, `RR2.Median`, `RR2.WMean`, `NPRP`, `NPRA`, `FVCI`, `FVCI.Mean`, `FVCI.Median`, `FVCI.WMean`
  - Transformed variable: `inc.l`
  - Dummy variables: `AA`, `educ.6`, `em.dur.5p`, `use.m`, `ver.2`, `Mining`, `Utilities`
  - For two highly correlated variables, keep one and delete the other one: `time2`, `time3`, `FreeCash.d`, `previous.loan.l`

Data Sampling: After data cleaning, we have 12228 positive (default) and 20241 negative (non-default) records. We randomly select 12000 positive and 12000 negative records to construct the sample.

An Overview on Features in Cleaned Dataset

After data cleaning, 155 features remain. Here we give an overview.

Group	Example	Explanation	Other variable in this group
Dependent variable	`Default`	1 - loan defaulted, 0 - otherwise	
Demographic information	`Age`	The age of the loan applicant	`Gender`, `Marital status`, `Kids`, ⋯
Borrower's financial information	`DebtToIncome`	Ratio of borrower's monthly gross income that goes toward paying loans	`Salary`, `Current debt`, ⋯
Relevant to the specific loan	`Loan amount`	Estimated amount the borrower has to pay every month	`Interest rate`, `Loan duration`, ⋯
Relevant to the application	`Hour of application`	Application hour	`Monday`, `After midnight`, ⋯

Notation

Notation	Explanation
`variable`	This font indicates variables in the initial dataset.
`variable`	This font indicates parameters in the supervised model.
(N, g)	The graph.
$N = \{1, 2, \cdots, i, j, \cdots, n\}$	The nodes set.
$g = \{ij : i, j \in N\}$	Edges in graph (N, g).
x_p	The pth feature in the cleaned dataset.
d_{ij}	The Gowers' distance between node i and node j.
w_{ij}	The weight of edge ij.

Build up the Graph

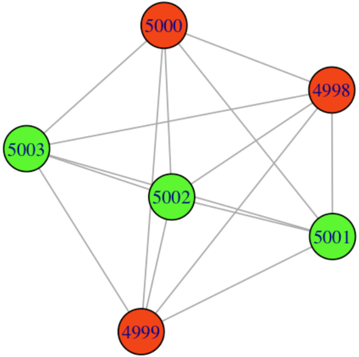


Figure 1: A Graph Example

Nodes and edges:

- Each node in the graph is a loan;
- We use Gower's distances [1] between two nodes as the weight of edge ij (w_{ij}).

$$d_{ij} = w_{ij} = \sum_{p=1}^{P} \frac{1}{P} \times \frac{d_{ij}^{p}}{\max(x_{\cdot p}) - \min(x_{\cdot p})},$$

where $\mathrm{d}_{ij}^{p} =$

$$\begin{cases} |x_{ip} - x_{jp}| & \text{if } x_p \text{ is a continuous variable,} \\ 1 & \text{if } x_p \text{ is a categorical variable and } x_{ip} \neq x_{jp}, \\ 0 & \text{if } x_p \text{ is a categorical variable and } x_{ip} = x_{jp}. \end{cases}$$

[1] Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. Biometrics, 27(4), 857–871. https://doi.org/10.2307/2528823

Minimum Spanning Tree (MST)

Definition

A Minimum Spanning Tree (MST) [1] is a subset of the edges of a connected, edge-weighted graph that connects all the vertices together, without any cycles and with the minimum possible total edge weight.

After the calculation of Gower's distance, we know the weights of all edges in a complete graph. We calculate the MST of this complete graph.

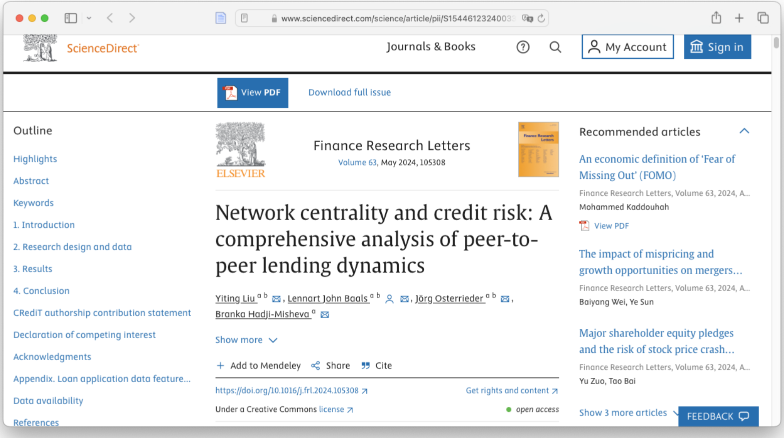
Reasons for using MST rather than complete graph:

- **Extract information efficiently:** On a complete graph, some centrality measures independent to edges' weights will be very close or identical. For example, degree centrality.
- **Keep a connected graph:** Centrality measures based on different component of the graph are not comparable.
- **Reduce computational cost:** Calculation of centrality measures are easier on MST.

[1] Prim, R. C. (1957). Shortest connection networks and some generalizations. The Bell System Technical Journal, 36(6), 1389-1401.

Table of Contents

Publication



Outline

- Highlights
- Abstract
- Keywords
- 1. Introduction
- 2. Research design and data
- 3. Results
- 4. Conclusion
- CRediT authorship contribution statement
- Declaration of competing interest
- Acknowledgments
- Appendix. Loan application data feature...
- Data availability
- References

Finance Research Letters
Volume 63, May 2024, 105308

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]

Show more

https://doi.org/10.1016/j.frl.2024.105308

Get rights and content

Under a Creative Commons license

open access

Recommended articles

An economic definition of 'Fear of Missing Out' (FOMO)
Finance Research Letters, Volume 63, 2024, A...
Mohammed Kaddouhah
View PDF

The impact of mispricing and growth opportunities on mergers...
Finance Research Letters, Volume 63, 2024, A...
Baiyang Wei, Ye Sun

Major shareholder equity pledges and the risk of stock price crash...
Finance Research Letters, Volume 63, 2024, A...
Yu Zuo, Tao Bai

Show 3 more articles

https://doi.org/10.1016/j.frl.2024.105308

Notation

Notation	Explanation
`variable`	This font indicates that the variable is in the original data set.
$G(V, E)$	A graph.
I	The total number of nodes.
$deg(i)$	The degree centrality of node i.
d_j	The distance between the ith node and the jth node.
$\mathrm{deg}^{\mathbf{neg}}(i)$	The non-defaulted group degree centrality for node i.
$\mathrm{deg}^{\mathbf{pos}}(i)$	The defaulted group degree centrality for node i.

Data Pre-Processing: from Initial Data to MST in Distance Matrix Form

Clean the data:

- Select loan records only in Estonian; Delete records with NA values;
- Drop columns:
  - Irrelevant variables: `date.start`, `date.end`
  - Variables cannot be known before a loan ends: `return`, `RR1`, `RR2.Mean`, `RR2.Median`, `RR2.WMean`, `NPRP`, `NPRA`, `FVCI`, `FVCI.Mean`, `FVCI.Median`, `FVCI.WMean`
  - Repeated income variables, as they express income in a transformed way: `inc.*.l`
  - Dummy variables: `AA`, `educ.6`, `em.dur.5p`, `use.m`, `ver.2`, `Mining`, `Utilities`
  - For two highly correlated variables, keep one and delete the other one: `time2`, `time3`, `FreeCash.d`, `previous.loan.l`

Data Sampling: After data cleaning, we have 12228 positive (defaulted) and 20241 negative (non-defaulted) records. We randomly select 6000 positive and 10000 negative records to construct the sample.

Calculate Gower's Distance: Gower's distance is not affected by different scales of variables, and is applicable to a combination of continuous data and categorical data.

Calculate Minimum Spanning Tree (MST) and express MST in distance matrix form: We use `Inf` in the matrix if two nodes do not have an edge in the MST.

Data Segment: We segment the data into training set and testing set.

Degree Centrality: Defaulted Group and Non-Defaulted group

Definition (Degree Centrality)

The Degree Centrality of a node i in a graph $G(V, E)$ is defined as:

$$\deg(i) = \sum_{j=1, j \neq i}^{I} Indicator[\text{node } j \text{ is connected to node } i]$$

What it measures: Degree Centrality quantifies the immediate influence or connectivity of a node within the network. A higher value indicates that the node is connected to a larger number of other nodes, thereby serving as a major junction or hub in the network. This metric is particularly useful for identifying nodes that serve as key connectors or influencers within the network topology.

Degree Centrality: Defaulted Group and Non-Defaulted Group

Operation: Delete the *i*th column

$$\text{ith row: } d_1, d_2, d_3 = \mathtt{Inf}, d_4 = \mathtt{Inf}, \cdots, d_i, \cdots, d_{J-2}, d_{J-1}, d_J$$

We delete the *i*th column because we do not consider the defaulted/non-defaulted status of the node itself.

Degree Centrality: Defaulted Group and Non-Defaulted Group

Operation: Delete columns not in the training set

$$\text{ith row: } d_1, d_2, d_3 = \texttt{Inf}, d_4 = \texttt{Inf}, \cdots, d_i, \cdots, d_{J-2}, d_{J-1}, d_J$$

We delete columns not in the training set because we do not know whether a point is defaulted or non-defaulted in the testing set.

Degree Centrality: Defaulted Group and Non-Defaulted Group

Operation: Delete `Inf` elements

$$\text{ith row: } d_1, d_2, d_3 = \texttt{Inf}, d_4 = \texttt{Inf}, \cdots, d_i, \cdots, d_{J-2}, d_{J-1}, d_J$$

We delete `Inf` elements because the node is not connected to these points in the MST.

Degree Centrality: Defaulted Group and Non-Defaulted Group

Operation: Count in the remaining columns

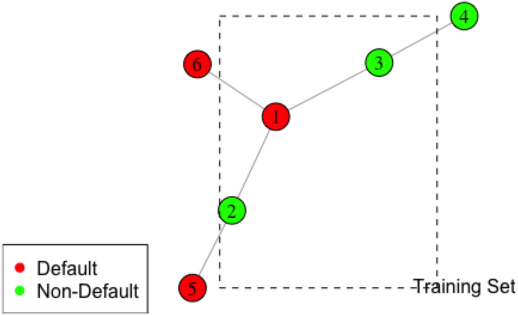
$$\text{ith row: } {\color{red}d_1}, {\color{green}d_2}, d_3 = \texttt{Inf}, d_4 = \texttt{Inf}, \cdots, d_i, \cdots, d_{J-2}, d_{J-1}, d_J$$

For the remaining columns, they indicate that the i-th node is connected to this in-training-set point in the MST. We count in the remaining columns, how many are matched with defaulted loans and how many are matched with non-defaulted loans.

We use $\deg^{\text{neg}}(i)$ to denote the non-defaulted group degree centrality for node i; we use $\deg^{\text{pos}}(i)$ to denote the defaulted group degree centrality for node i.

Degree Centrality: Defaulted Group and Non-Defaulted Group

An example



Node	deg^{pos}	deg^{neg}
1	0	2
2	1	0
3	1	0
4	0	1
5	0	1

Logistic Regression and Performance Analysis

We add $\deg^{\text{neg}}(v)$ and $\deg^{\text{pos}}(v)$ back to initial data set as two additional graph features;

We normalize the data by subtracting the mean and dividing by the standard deviation;

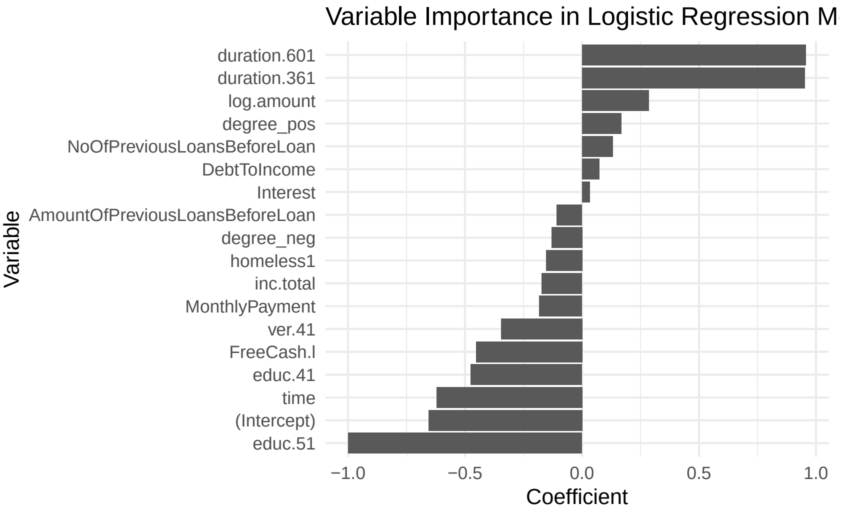
We identify the top 20 features based on their Mean Decrease Gini scores in the Random Forest model;

We run the logistics regression with step wise. This step continues to select features during regression.

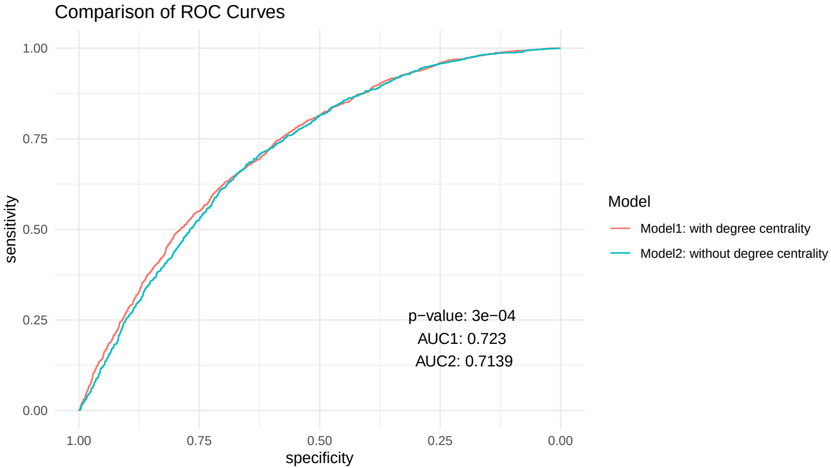
We conduct DeLong test to compare the regression model based on data with two extra degree centrality measures and without.

We shuffle the two extra degree centrality measures and fit the model again. If the improvement disappears, it means our improvement in regression model brought by two extra degree centrality measures is robust.

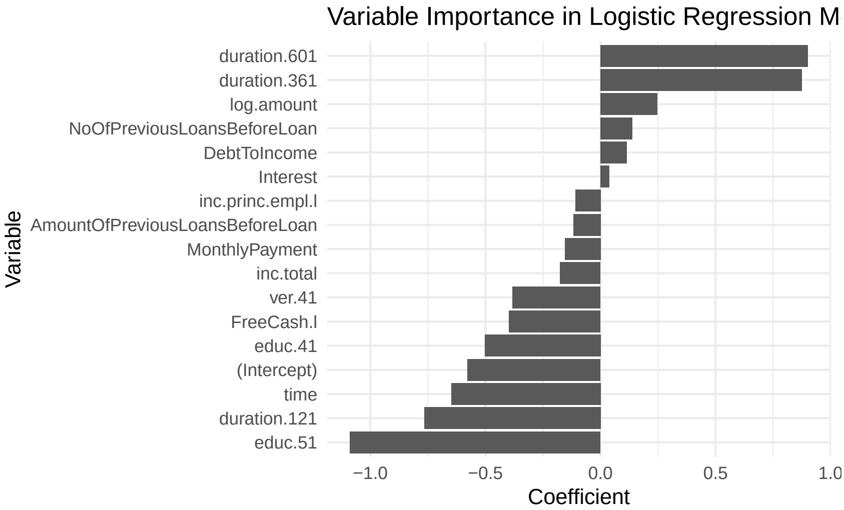
Feature Importance



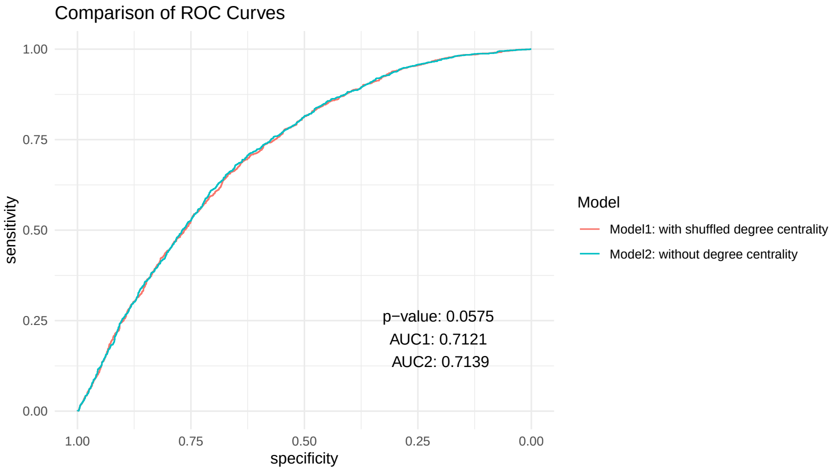
ROC and DeLong Test



Robustness Check - Feature Impotence



Robustness Check - ROC and DeLong Test


Comparison of ROC Curves
sensitivity
specificity
p−value: 0.0575
AUC1: 0.7121
AUC2: 0.7139
Model
Model1: with shuffled degree centrality
Model2: without degree centrality