Network-based scoring models to improve credit risk management in peer to peer lending platforms

Branka Hadji Misheva, ZHAW
Paolo Giudici, UNIPV
Alessandro Spelta, UNIPV

FIN-TECH HO2020 project

Zürcher Hochschule für Angewandte Wissenschaften

zhaw

FINTECH RISK MANAGEMENT
www.fintech-ho2020.eu

European Commission

UCD DUBLIN
UNIVERSITY OF TAMPERE
UNIVERSITÀ DI PAVIA
Jožef Stefan Institute
UCL
uni.lu UNIVERSITÉ DU LUXEMBOURG
WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS
ktu faculty of mathematics and natural sciences 1922
INESCTEC
UNIVERSIDAD COMPLUTENSE MADRID
HUMBOLDT-UNIVERSITÄT ZU BERLIN
POLITECNICO MILANO 1863
PANTEION UNIVERSITY OF SOCIAL AND POLITICAL STUDIES
MUNI
UNIVERSITÉ PARIS 1 PANTHÉON SORBONNE
UNIVERSITY OF WARSAW

Structure

Section 1
Description of Use Case – Using Network Models for the purpose of improving predictive utility of scoring models in P2P systems

Section 2
Data and Methodology

Section 3
Results and Comparison of Predictive Utility

Section 4
Lets see it!!

Section 1
Description of Use Case

Fintech overview

- FinTech (i.e. financial technology) denotes companies that combine financial services with modern innovative technologies.

- The advances in IT has enabled online markets to provide an alternative to traditional financial intermediaries.

- **With the increasing role of these online lending marketplaces, key point of interest becomes assessing the risk associated with these new players.**

Focus of the study

- We focus on fintech credit.

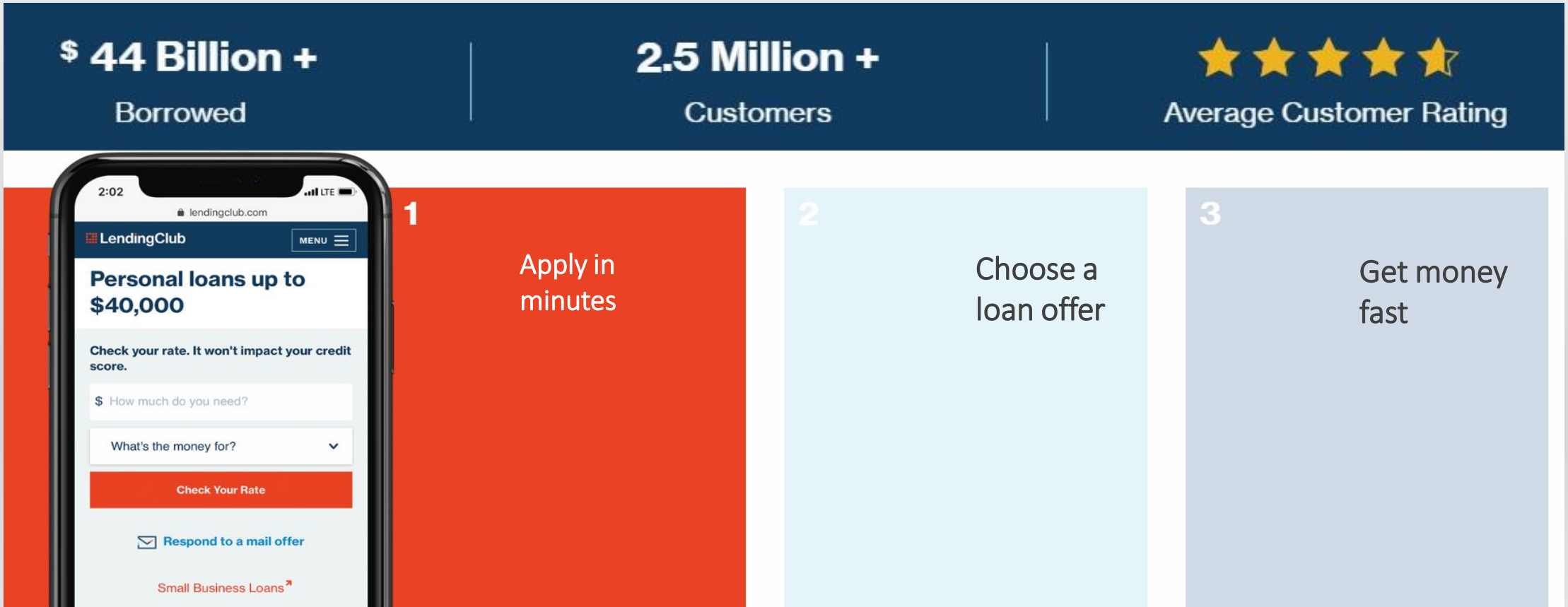
- Financial Stability Board: *"Fintech credit - all credit activity facilitated by platforms that match borrowers with lenders"*.

- Other terms:
  - P2P Lending Platforms
  - Loan-based crowdfunders
  - Marketplace lenders

- These online platforms allow lenders and borrowers to match themselves and furthermore, provide a rating of borrowers' creditworthiness.

P2P Lending: Importance

- Small % of overall credit but its growing rapidly;
- In the US, 36% of unsecured personal lending was issued by FinTech in 2017;
- LendingClub, the world's largest P2P lending platform, in 2019 has reached more than $44billion in total loan issuance.



Rapid Growth

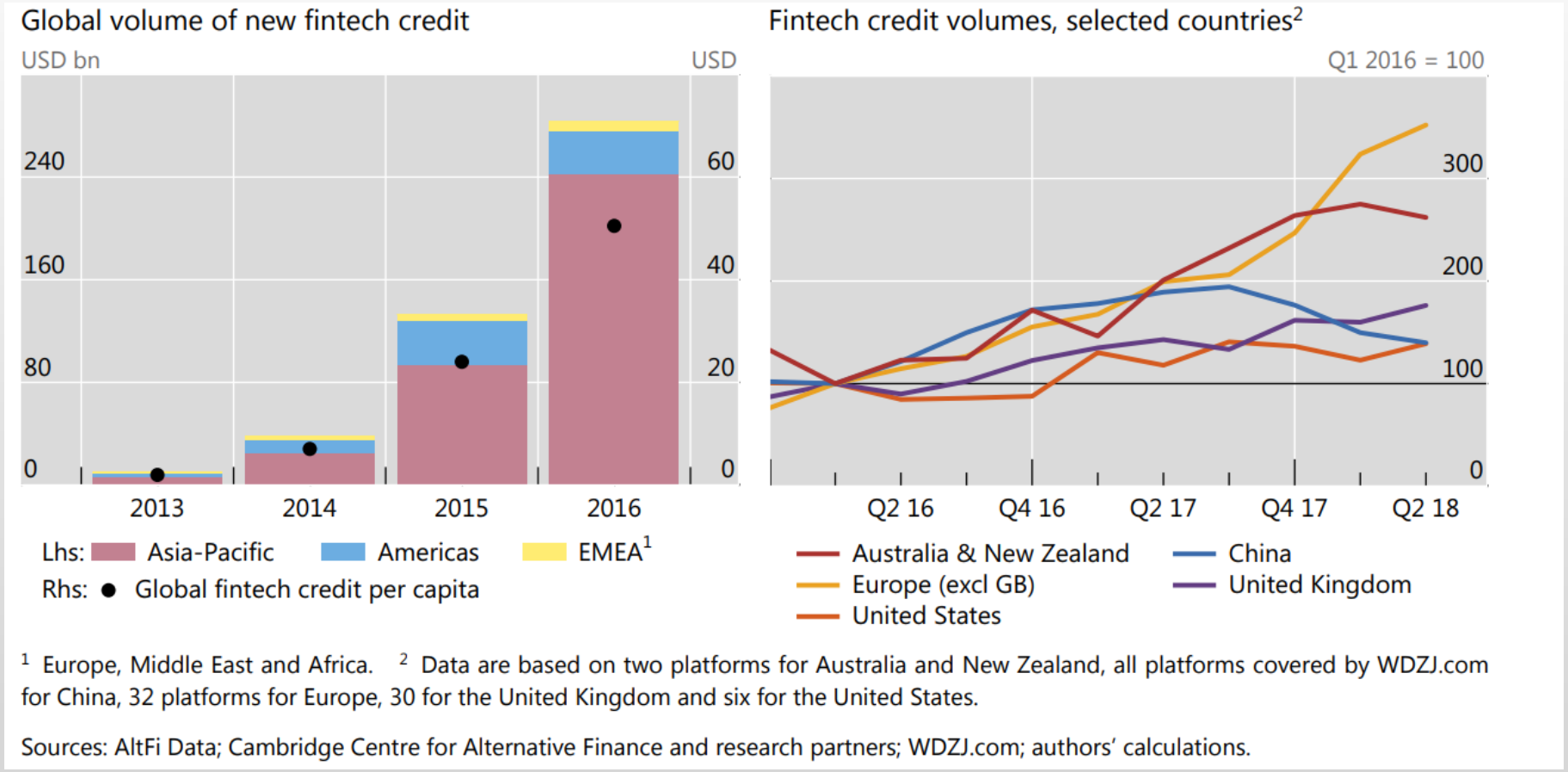


Figure 1. Growth of FinTech (alternative) Credit
Source: https://www.bis.org/publ/qtrpdf/r_qt1809e.pdf

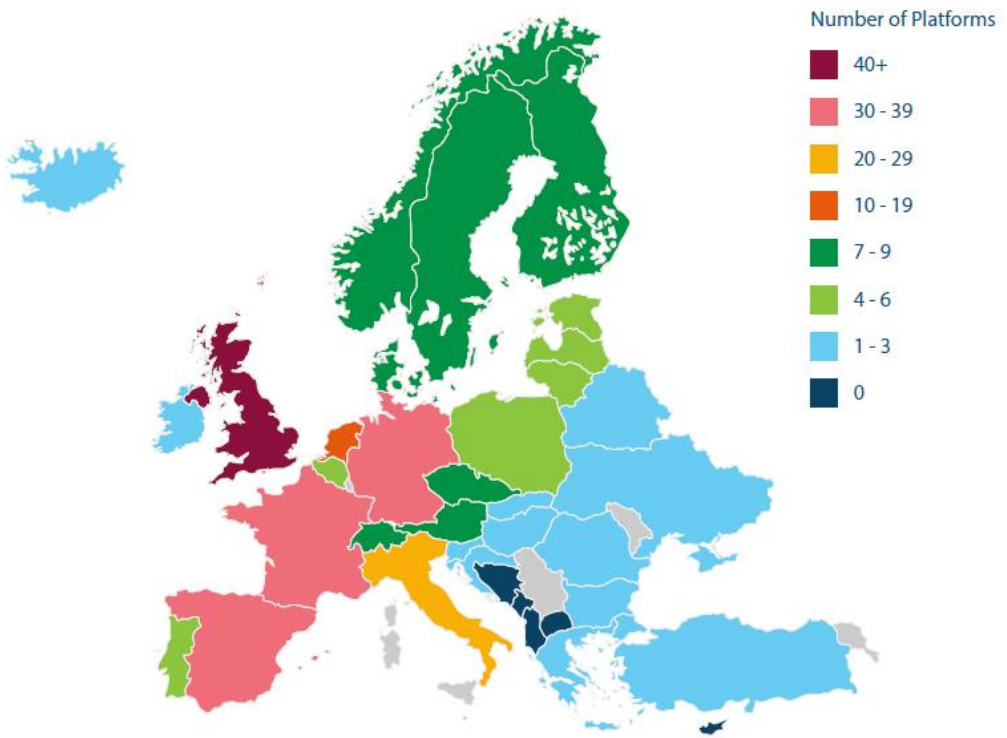


Figure 2. Number of platforms per region
Source: Cambridge Center for Alternative Finance

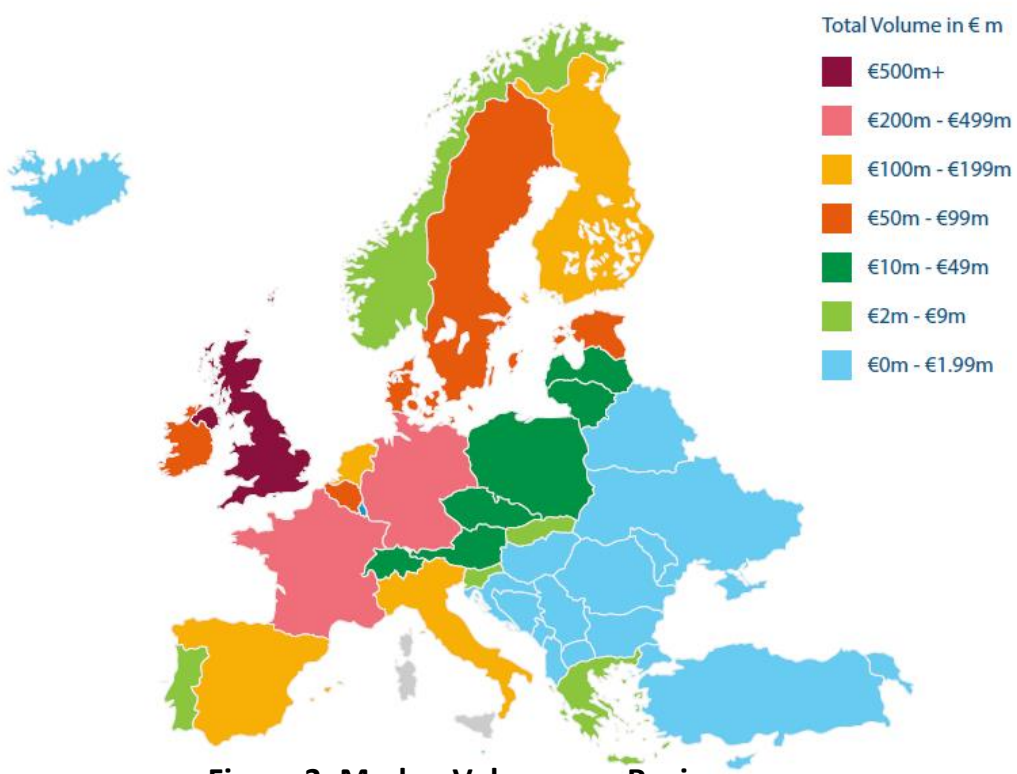


Figure 3. Marker Volume per Region
Source: Cambridge Center for Alternative Finance

Geographic Distribution and Market Volume

Advantages associated with P2P lending platforms

- Can avoid many intermediation costs;
- Use of non-traditional data source
  - Greater convenience –> Fuster et al (2018) find that fintech lenders in the United States improve borrower convenience by processing mortgages 15–30% faster than other lenders, on average;
  - Berg et al (2018) find that a German P2P platform's default risk assessments, incorporating data on the digital footprints of customers registered on its website, outperform its assessments based on credit bureau data alone.
  - Widened access to credit –> De Roure et al (2016), Jagtiani and Lemieux (2018a), Baeck et al (2014), US Department of the Treasury (2016).

Disadvantages associated with P2P lending platforms

The advantages notwithstanding, the growth of alternative financial institutions is associated with several causes for concern:

- Less able to deal with asymmetric information;
- Cannot sustain the cost of monitoring the clients once a loan has been assigned;
- Investors are not protected (in most cases) in case of failure;
- Difference in risk ownership.

Difference in risk ownership

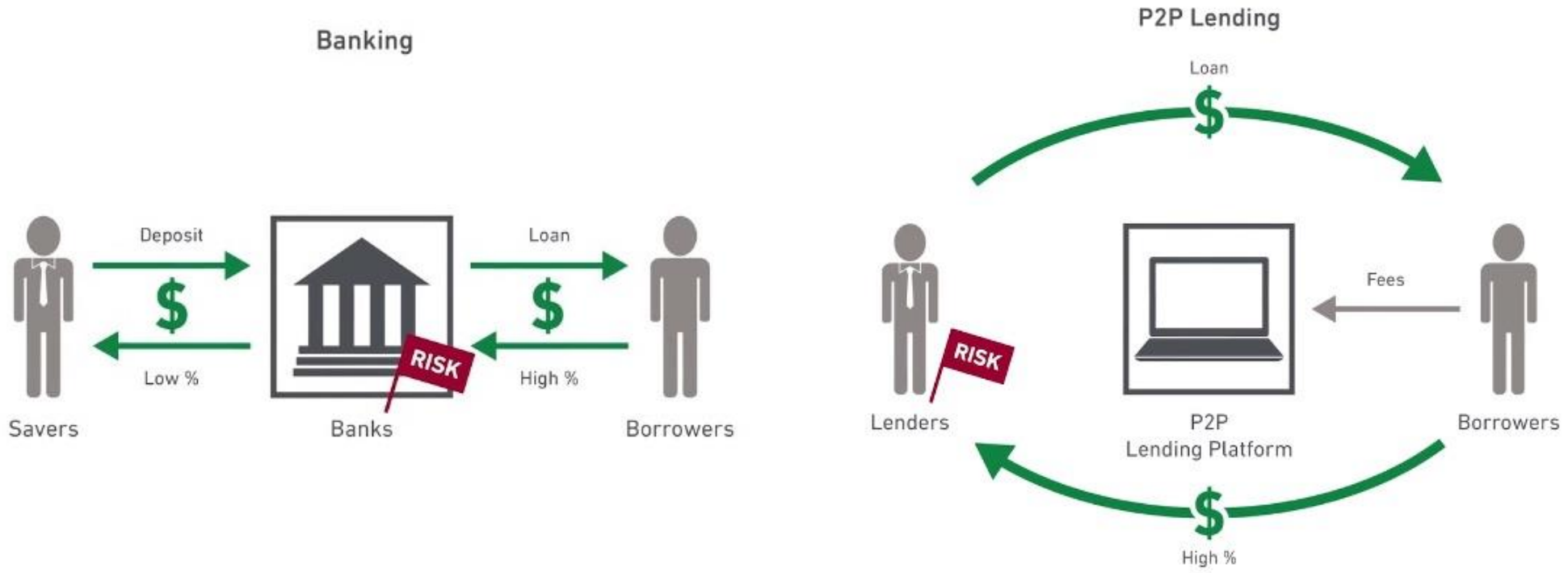


Figure 4: Difference is risk ownership (Banking vs P2P Platforms)

Objectives of the study

To cope with these issues, we propose **to exploit the topological information embedded into similarity networks generated by P2P participants.**

Specific objectives of the use case:

- To analyze the predictive performance of scoring models employed by P2P platforms, specializing in SME lending;
- To investigate whether network information can improve loan default predictions and further protect lenders, in a financial stability context.

Section 2
Data and Methodology

Data and Methodology

- Data is collected from a European Fintech specializing in financial consultancy and evaluation of companies creditworthiness.

- Specifically, the analysis relies on financial data on 15,000 SMEs which are the target of P2P lending platforms.

- The proportion of defaults in the sample is equal to 11%.

Data Used

ID	FORMULA	ID	FORMULA
RATIO001	(Total assets - Shareholders Funds)/Shareholders Funds	RATIO019	Interest paid/(Profit before taxes + Interest paid)
RATIO002	(Long term debt + Loans)/S. Funds	RATIO027	EBITDA/interest paid
RATIO003	Total assets/Total liabilties	RATIO029	EBITDA/Operating revenues
RATIO004	Current assets/Current liabilties	RATIO030	EBITDA/Sales
RATIO005	(Current assets - Current assets: stocks)/Current liabilties	RATIO036	Constraint EBIT
RATIO006	(Shareholders Funds + Non current liabilities)/Fixed assets	RATIO037	Constraint PL before tax
RATIO008	EBIT/interest paid	RATIO039	Constraint Financial PL
RATIO011	(Profit (loss) before tax + Interest paid)/Total assets	RATIO040	Constraint P/L for period th EUR
RATIO012	P/L after tax/Shareholders Funds	DPO	Trade Payables/Operating revenues
RATIO013	GROSS PROFIT/Operating revenues	DSO	Trade Receivables/Operating revenues
RATIO017	Operating revenues/Total assets	DIO	Inventories/Operating revenues
RATIO018	Sales/Total assets	NACE	Industry classification on NACE

Table 1. Description of variables included in the dataset.

Credit Risk Models

- Credit risk models are useful tools for modeling and predicting individual firm's default;
- Regression techniques or machine learning approaches;
- Consider N firms having observation regarding T different. For each institution n define a variable Yn to indicate whether such institution has defaulted on its loans or not, i.e. $Yn = 1$ if company defaults $Yn = 0$, otherwise;
- In a nutshell, credit risk models develop relationships between the explanatory variables embedded in T and the dependent variable Y.

Classifiers

- **Logistic Regression** → The logistic regression has been one of the most widely used methods for evaluating the probability of default of an entity

- **Linear Discriminant Analysis (LDA)** → This model assumes that different classes generate data based on different Gaussian distributions. The method approaches the problem by assuming that the conditional probability density functions $p(x\,|y=0)$ and $p(x\,|y=1)$ are both normally distributed with mean and covariance parameters (μ_0, V_0) and (μ_0, V_0) respectively.

- **Classification and Regression Trees (CART)** → Another widely used statistical technique in which a dependent variable is associated with a set of input factors through a recursive sequence of simple binary relations.

- **Support Vector Machine (SVM)** → This approach classifies data by detecting the best hyperplane that separates all data points of one class from those of the other class.

Assessing Model Performance

- Receiver operating characteristic (ROC) curve

$$FPR = \frac{FP}{FP+TN} \quad and \quad TPR = \frac{TP}{TP+FN}$$

- Overall accuracy

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

- KS statistic

$$KS = max_j \left| F_{Active}(x_j) - F_{Defaulted}(x_j) \right|$$

- Back-testing: sub-sampling validation approach. The results concerning the model accuracy (area under the ROC curve, KS statistic, Gini index) are then averaged over the splits.

Distance Metric

- We exploit information derived from financial statements of borrowing companies collected in a vector representing the composition of the financial ratios of institution n.

- We define a metric $d_{x,y}$ that provides the relative distance between companies by applying the standardized Euclidean distance between each pair of institutions feature vectors.

$$d_{x,y} = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}$$

- Namely, each coordinate difference between pairs of vectors (xj - yj) is scaled by dividing by the corresponding element of the standard deviation.

The Minimal Spanning Tree

- $d_{x,y}$ is fully connected → does not help to find out whether there exist dominant patterns of similarities between institutions;

- Therefore, we derive the Minimal Spanning Tree (MST) representation of borrowing companies' financial ratios' similarities;

- For a Graph G, the goal is to find a tree T which is a spanning subgraph of G, i.e. every node is included to at least one edge of T, and has minimum total weight:

  - Pick some arbitrary start node u. Initialize T = {u};
  - At each step add the lowest-weight edge to T(the lowest-weight edge that has exactly one node in T and one node not in T);
  - Stop when T spans all the nodes.

Network Measures and Community Detection

- For each node (borrower company), we compute the **degree** and **strength centrality**

- The degree of a vertex - the number of edges incident to it.

$$\widehat{D}_{ij} = \begin{cases} 1 \; if \; d_{ij} > 0 \\ 0 \; otherwise \end{cases}$$

$$K_i = \sum_{j=1}^{N} \widehat{D}_{ij}$$

- Similarly, the strength centrality measures the average distance of a node with respect to its neighbours.

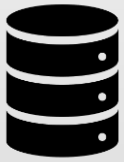
$$s_i = \sum_{j=1}^{N} D_{ij}$$

- We also apply the Louvain Method to extract the community structure of the network.

Description of procedure

Input Data

Input Data



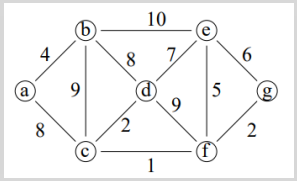
Euclidean Distance

$$d_{x,y} = \sqrt{\sum_{j=1}^{J} (x_j - y_j)^2}$$

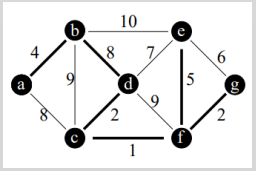
Adjacency matrix

$$\begin{pmatrix} 0 & 4 & 8 & 0 & 0 & 0 & 0 \\ 4 & 0 & 9 & 8 & 10 & 0 & 0 \\ 8 & 9 & 0 & 2 & 0 & 1 & 0 \\ 0 & 8 & 2 & 0 & 7 & 9 & 0 \\ 0 & 10 & 0 & 7 & 0 & 5 & 6 \\ 0 & 0 & 1 & 9 & 5 & 0 & 2 \\ 0 & 0 & 0 & 0 & 6 & 2 & 0 \end{pmatrix}$$

Graph



MST



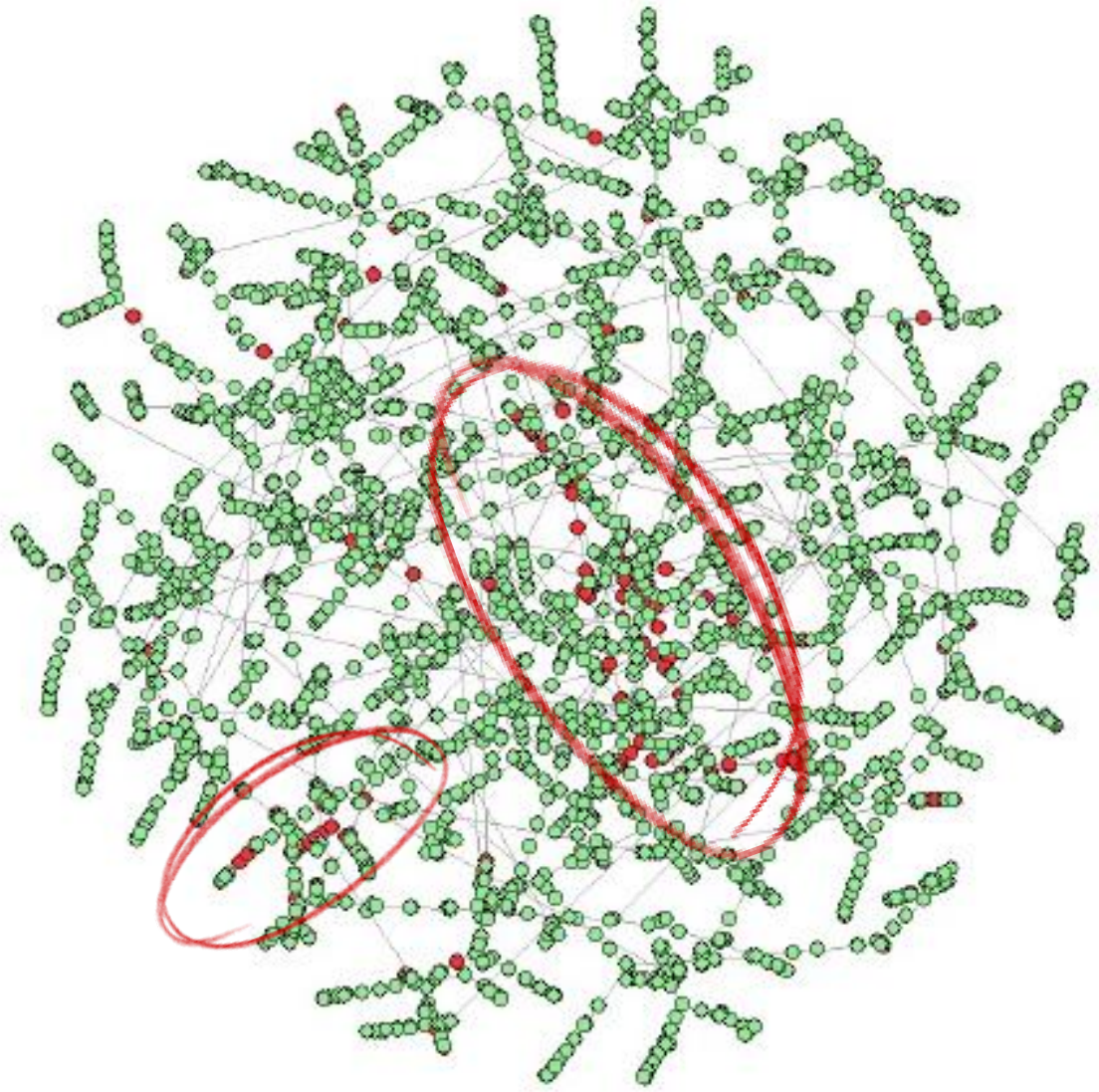
Centrality and community

Input Data + Network Parameters

Section 3
Results and Comparison of Predictive Utility

Data Used: Summary Statistics

Figure 5. Minimal spanning tree representation of the borrowing companies networks. In the panel, nodes are colored according to their financial soundness, red nodes represent defaulted institutions while green nodes are associated with active companies



Results: Predictive accuracy comparison

	AUC		KS		Gini		Accuracy	
	Basic	Network	Basic	Network	Basic	Network	Basic	Network
Logit	79.631	80.793	52	52	59.262	61.586	90.193	90.09661
LDA	77.759	79.16	51	52.8	55.518	58.32	90.122	89.98844
CART	67.973	67.973	35.5	35.946	35.946	35.5	90.832	90.82413
SVM	76.81	77.65	53.62	50	51	55.3	92.44444	92.22222

Table 2. Summary statistics of the non-parametric analysis. From the left to the right: area under the ROC curve (AUC), KS Statistic (KS), Gini Index (Gini) and the model accuracy (Accuracy). For each measure and for all the tested models, we report the results obtained by the baseline scenario and for the network-augmented configuration

	AUC		AUPR		ACC	
	Basic	Network	Basic	Network	Basic	Network
LOGIT	0.7252	0.8021	0.1827	0.2653	0.9376	0.951
LDA	0.7197	0.7197	0.259	0.2766	0.9443	0.9376
SVM	0.6014	0.716	0.1361	0.1556	0.9398	0.942
CART	0.716	0.7178	0.234	0.2416	0.9354	0.9376

Table 3. Summary statistics of the non-parametric analysis. From the left to the right: area under the ROC curve (AUC), area under the PR curve (AUPR) and model accuracy (ACC). For each measure and for all the tested models, we report the results obtained by the baseline scenario and for the network-augmented configuration

Results: Predictive accuracy comparison

Results: Predictive accuracy comparison (robustness check)

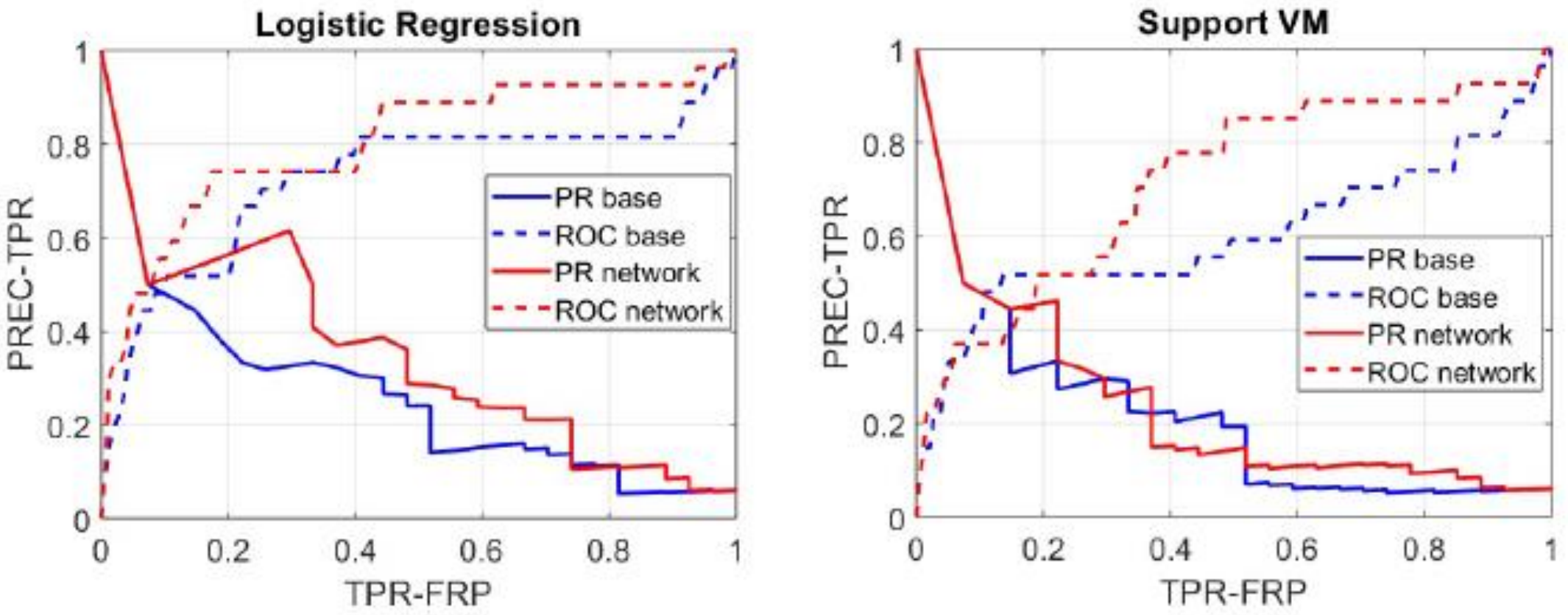


Figure 6. Precision Recall (PR) and Receiver Operating Characteristic (ROC) curves for the baseline credit risk models and for the network-augmented models. In each panel, dotted lines represent the ROC curves while solid lines refer to PR curves. In blue, we show the results related to the baseline models while in red we show the results related to the network-augmented models.

Conclusion

- We try to investigate whether topological information embedded into similarity networks generated by P2P participants can improve predictive accuracy of scoring models;
- Our results are promising ... however, how we define the network is of crucial importance;
- Next steps: financial flows between borrowers;
- The paper, data set used as well as the code to replicate the use case is available on our fintech-ho2020 platform.

Section 4
Lets see it!!