FINTECH RISK MANAGEMENT
www.fintech-ho2020.eu

REGTECH WORKSHOP I

Peer to Peer lending risk management

Branka Hadji Misheva, ZHAW Zurich University of Applied Sciences

modefinance

zhaw

European Commission

FINTECH OVERVIEW

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

- The advances in IT have 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 Peer to Peer (P2P) lending platforms, which allow private individuals to make small, unsecured loans to other private borrowers or small companies.

P2P platforms allow lenders and borrowers to match themselves, **providing a rating of borrowers' creditworthiness**.

advantages

The advantages associated with P2P lending platforms

- Factors that explain the increasing role of alternative financial institutions in the global world of finance:

- Permit to avoid many intermediation costs;
- Offer better interest rates to borrowers and higher rates of return for lenders;
- Use of big data analytics.

disadvantages

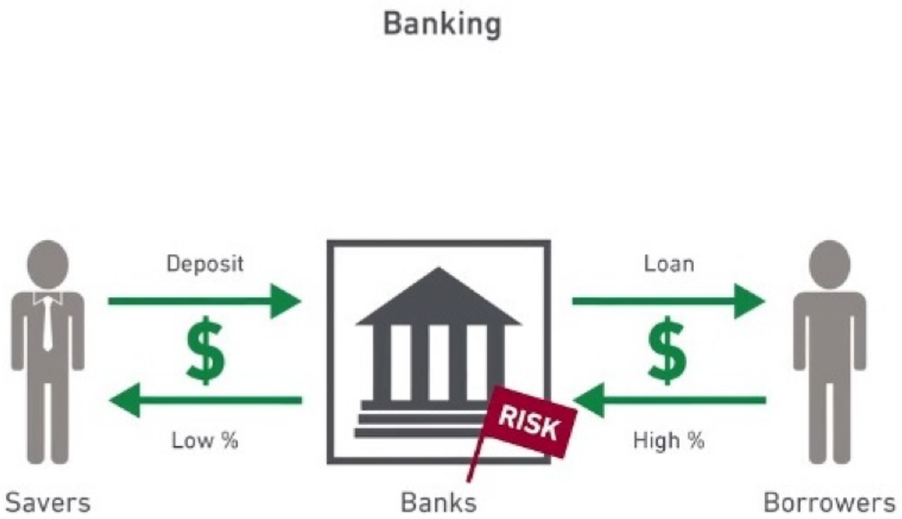
The disadvantages associated with P2P lending platforms

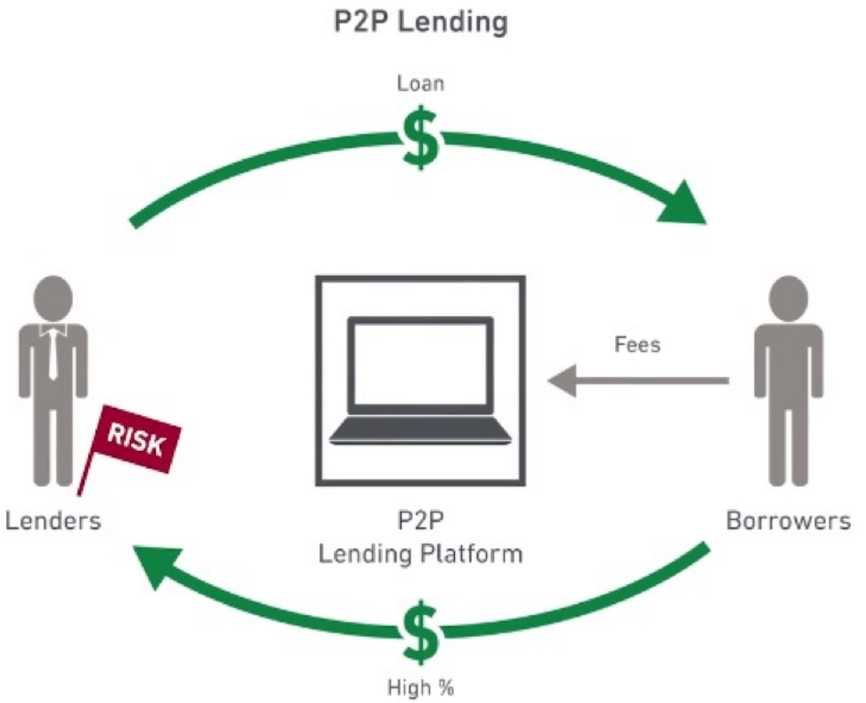
- The advantages notwithstanding, the growth of alternative financial institutions is associated with several causes for concern:

- Less able to deal with asymmetric information;
- Cannot sustain the cost of monitoring the clients once a loan has been assigned;
- Difference in risk ownership;
- No regulation (yet).

difference in risk ownership

Banking VS P2P lending





OBJECTIVES OF THE STUDY

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

- 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.

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 **4514 SMEs** which are the target of P2P lending platforms.

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

data used: summary statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max	Active	Default
ratio001	4,514	8.885	19.155	−64.430	1.303	9.680	206.550	8.85	9.15
ratio002	4,514	1.264	3.333	−10	0	1.2	33	1.25	1.35
ratio003	4,514	1.444	0.761	0.170	1.070	1.520	8.270	1.49	1.09
ratio004	4,514	1.536	1.201	0.010	0.970	1.720	13.710	1.6	1.04
ratio005	4,514	1.190	1.024	0.000	0.610	1.407	10.880	1.24	0.76
ratio006	4,514	7.726	23.277	−33.140	0.940	4.890	297.020	7.93	6.09
ratio008	4,514	23.068	70.271	−285.860	1.240	16.317	566.960	26.22	-2.33
ratio011	4,514	0.028	0.147	−1	0.01	0.1	0	0.05	-0.13
ratio012	4,514	−0.069	0.790	−8.540	0.000	0.210	1.080	0.01	-0.69
ratio017	4,514	1.372	1.068	0.010	0.680	1.740	8.420	1.38	1.30
ratio018	4,514	1.335	1.064	0.010	0.640	1.700	8.420	1.34	1.29
ratio019	4,514	0.194	0.498	−3.320	0.010	0.390	3.950	0.21	0.05
ratio027	4,514	36.513	92.893	−191.630	2.470	27.608	747.010	40.18	6.96
ratio029	4,514	0.062	0.196	−2	0.02	0.1	1	0.08	-0.12
ratio030	4,514	0.068	0.216	−2	0.02	0.1	1	0.09	-0.12
DIO	4,514	105.228	355.807	0	1	80	5,569	100.61	142.47
DPO	4,514	75.934	111.651	0	0	99.8	1,467	67.35	145.18
DSO	4,514	95.732	128.370	0	0	136	1,465	91.07	133.32
turnover	4,514	3,344.479	7,580.559	6	594	2,761.8	76,403	3542.27	1749.41

Summary statistics of variables included in the dataset.
For each measure we report the average (Mean) along with the standard deviation (St. Dev.), the minimum (Min), the 25-th and 75-th percentiles (Pctl), the maximum (Max), mean value of the variable for active companies (Active), mean value of the variable for defaulted companies (Defaulted)

CREDIT RISK MODELS

Credit risk models are useful tools for modeling and predicting individual firm's default.

Such models are usually grounded on regression techniques or machine learning approaches often employed for financial analysis and decision-making tasks.

Consider *N* firms having observation regarding *T* different variables (usually balance-sheet measures or financial ratios). For each institution *n* define a variable γ_n to indicate whether such institution has defaulted on its loans or not, i.e. $\gamma_n = 1$ if company defaults, $\gamma_n = 0$ otherwise.

In a nutshell, credit risk models develop relationships between the explanatory variables embedded in *T* and the dependent variable γ.

logistic regression

The **logistic regression** has been one of the most widely used methods to evaluate the **probability of default** of an entity.

$$ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \sum_j \beta_j x_{ij}$$

where *pi* is the probability of default, for borrower *i*, $x_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{iJ})$ is a vector of borrower-specific explanatory variables, and the intercept parameter α, as well as the regression coefficients β_j, for $j = 1, \ldots, J$, are to be estimated from the available data.

It follows that the probability of default can be obtained as:

$$p_i = \frac{1}{1 + e^{\alpha + \sum_j \beta_j x_{ij}}}$$

other classifiers

- 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 $p(\mathbf{x}|\gamma = 0)$ and $p(\mathbf{x}|\gamma = 1)$ are both normally distributed with mean and covariance parameters $(\mu_0, \mathbf{V}_0)$ and $(\mu_1, \mathbf{V}_1)$ 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:

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

- Overall accuracy:

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

- KS statistic:

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

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

distance metric

We exploit information derived from financial statements of borrowing companies collected in a vector $\mathbf{x}_n$ representing the financial composition of the balance-sheet of institution *n*.

We define a metric *D* that provides the relative distance between companies by applying the standardized Euclidean distance between each pair $(\mathbf{x}_i, \mathbf{x}_j)$ of institutions feature vectors.

$$d_{i,j} = (\mathbf{x}_i - \mathbf{x}_j)\triangle^{-1}(\mathbf{x}_i - \mathbf{x}_j)'$$

Namely, each coordinate difference between pairs of vectors $(\mathbf{x}_i - \mathbf{x}_j)$ is scaled by dividing by the corresponding element of the standard deviation.

Minimal Spanning Tree

Although *D* can be informative about the distribution of the distances between the companies, the fully-connected nature of this set 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' balance-sheet 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 (firm), we compute the degree and strength centrality.

- The degree k_i of a vertex *i* with $(i = 1, \ldots, N)$ is the number of edges incident to it.

$$\hat{\mathbf{D}}_{ij} = \begin{cases} 1 & if \quad d_{ij} > 0 \\ 0 & otherwise \end{cases}$$

then, the degree of a vertex *i* is:

$$k_i = \sum_{j=1}^{N} \hat{\mathbf{D}}_{ij}.$$

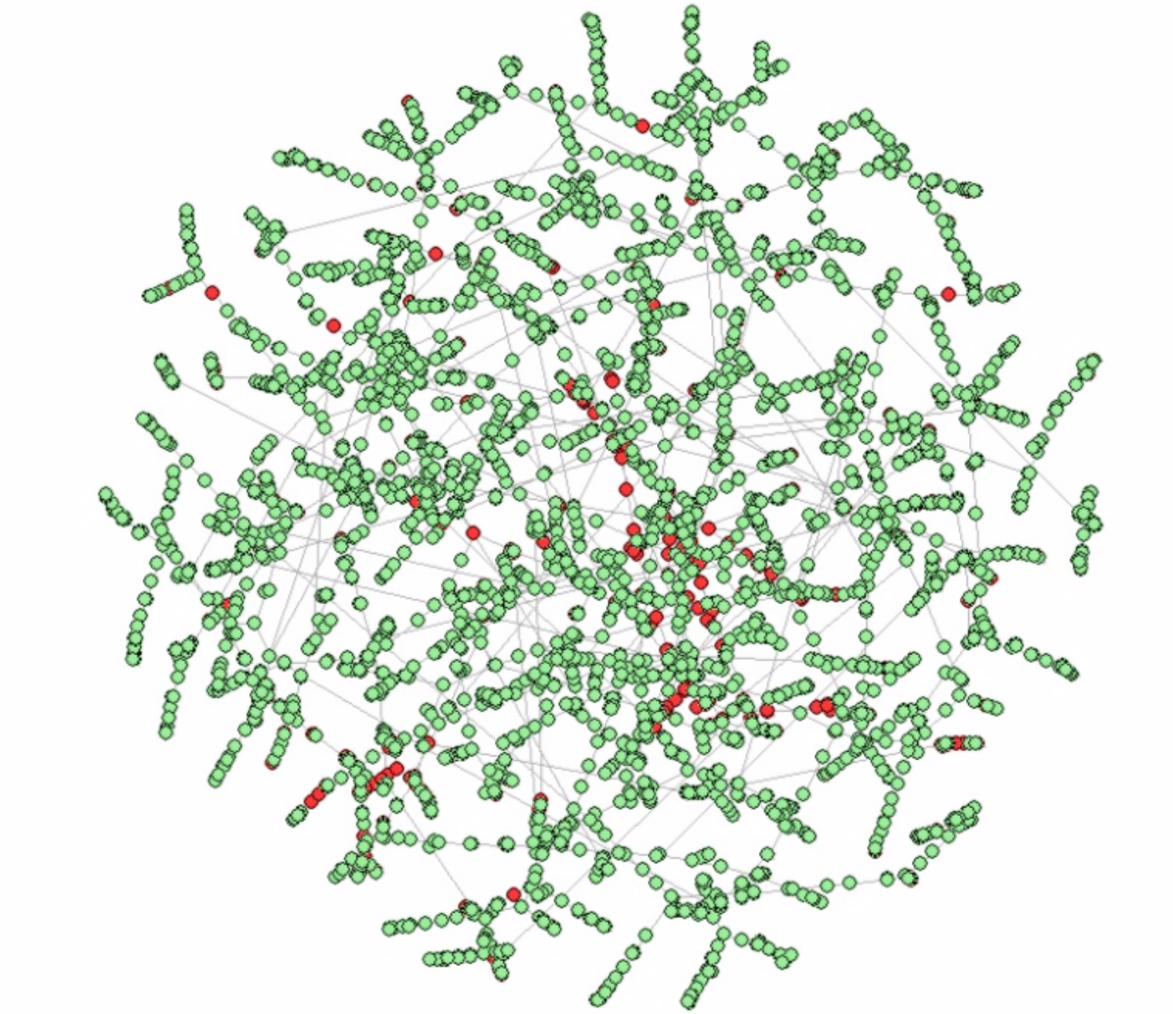
Similarly, the strength centrality measures the average distance of a node with respect to its neighbors.

$$s_i = \sum_{j=1}^{N} \mathbf{D}_{ij}.$$

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

RESULTS

MST representation of the network



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

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

RESULTS

Predictive accuracy comparison (robustness check)

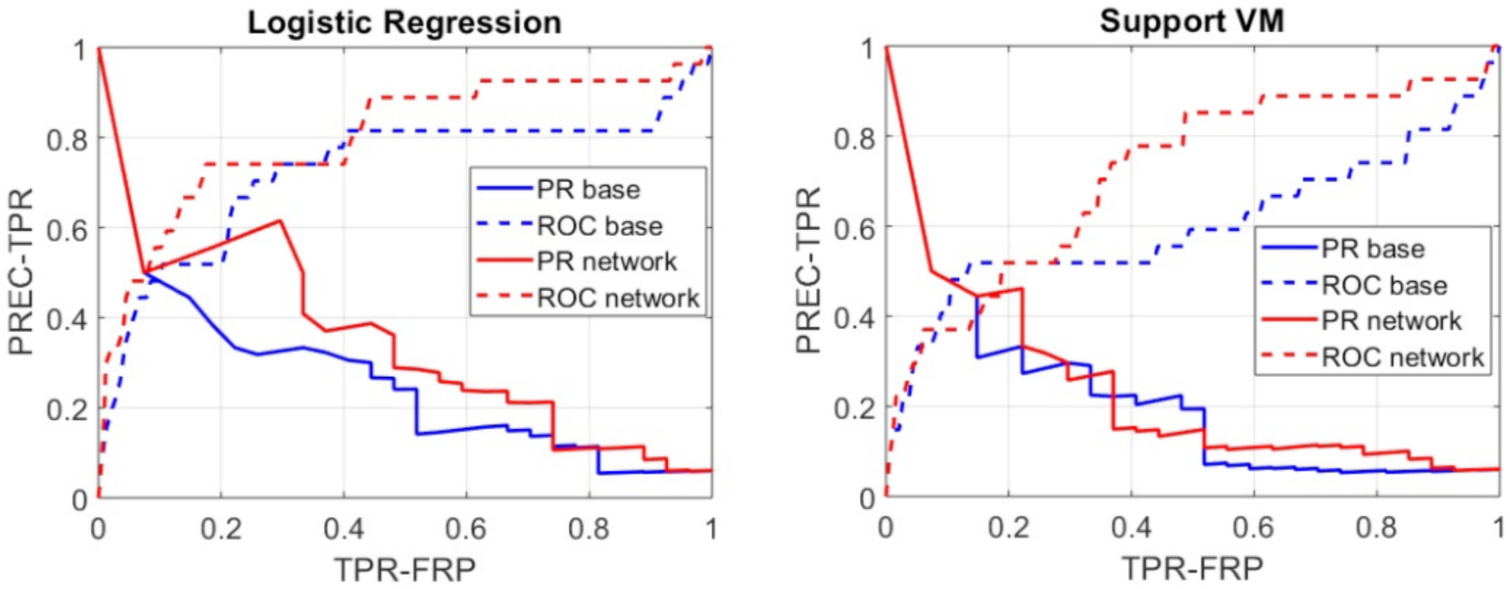
We ran the same analysis on a smaller data set with fewer financial ratios:

	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

From the left to the right: area under the ROC curve (AUC), area under the PR curve (AUPR), model accuracy (ACC).

RESULTS

Predictive accuracy comparison (robustness check)



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.

LET'S SEE THIS !