P2P Lending Systems

FinTech-ho2020

Winterthur, 3 September 2019

Use Case I: Network-based scoring models to improve creditrisk management in peer to peer lending platforms

Paolo Giudici, Branka Hadji Misheva and Alessandro Spelta

University of Pavia and ZHAW

Pros and Cons: P2P lending platforms

Pros:

- Can avoid many intermediation costs;
- Use of non-traditional data source;
- Greater convenience;
- Widened access to credit - De Roure et al (2016), Jagtiani and Lemieux (2018a), Baeck et al (2014), US Department of the Treasury (2016).

Cons:

- 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

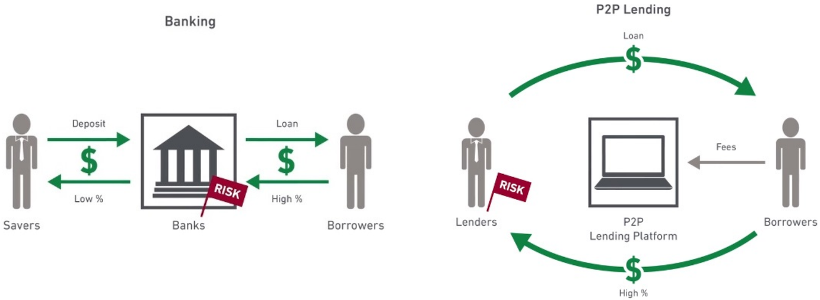


Figure: Difference is risk ownership (Banking vs P2P Platforms)

Objectives of the study

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

- Data is collected from a European Fintech specializing in financial consultancy and evaluation of companies' creditworthiness.
- Specifically, the analysis relies on financial data on 15,015 SMEs which are the target of P2P lending platforms.
- The proportion of defaults in the sample is equal to 11%.

Classifiers

- Logistic Regression (LR) - One of the most widely used methods for evaluating the probability of default of an entity.
- Linear Discriminat Analysis (LDA) - This approach models the distribution of predictors separately in each of the response classes, and then it uses Bayes theorem to estimate the probability.
- 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

$$FPR = \frac{FP}{FP + TN} \qquad \text{and} \qquad TPR = \frac{TP}{TP + FN} \tag{1}$$

- Overall accuracy

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

- KS statistic

$$KS = max_j |F_{Active}(x_j) - F_{Defaulted}(x_j)| \tag{3}$$

- 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 and MST

- We define a metric **D** - relative distance between companies;

$$D_{x,y} = \sqrt{\sum_{j=1}^{J} \left(\frac{x_j}{s_j} - \frac{y_j}{s_j} \right)^2} \tag{4}$$

- **D** is fully-connected
- We derive the Minimal Spanning Tree (MST) representation of borrowing companies' financial 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.

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 $(i = 1, ..., N)$ 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 a vertex i is:

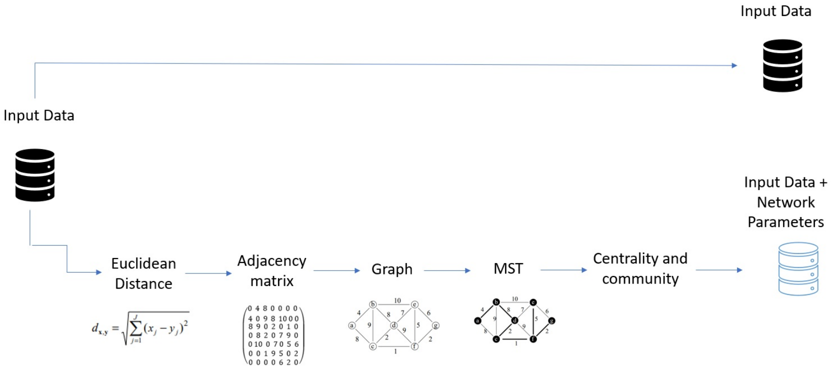
    $$k_i = \sum_{j=1}^{N} \hat{\mathbf{D}}_{ij}. \tag{5}$$

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

    $$s_i = \sum_{j=1}^{N} \mathbf{D}_{ij}. \tag{6}$$

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

Summary of Approach - Data



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

Results: MST representation of the network

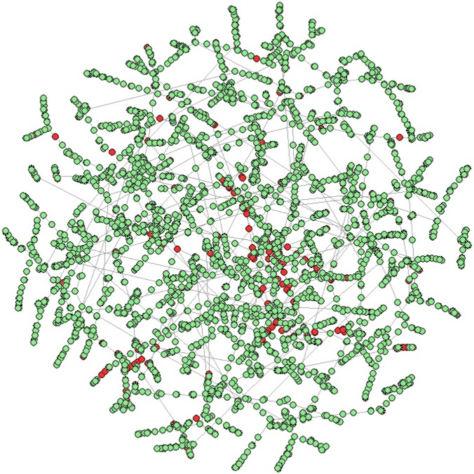
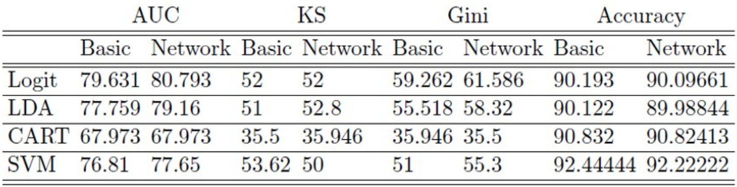


Figure: Minimal spanning tree representation of the borrowing companies networks

Results: Predictive accuracy comparison

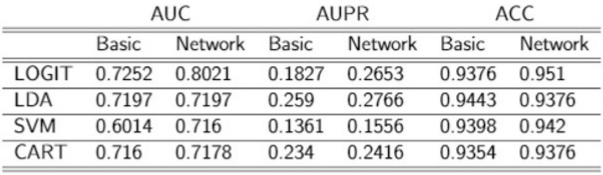


	AUC		KS		Gini		Accuracy	
	Basic	Network	Basic	Network	Basic	Network	Basic	Network
Logit	79.631	80.793	52	52	59.262	61.586	90.193	90.09661
LDA	77.759	79.16	51	52.8	55.518	58.32	90.122	89.98844
CART	67.973	67.973	35.5	35.946	35.946	35.5	90.832	90.82413
SVM	76.81	77.65	53.62	50	51	55.3	92.44444	92.22222

Table: Summary Statistics of non-parametric analysis. 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), 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

Table: 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), model accuracy (ACC)

Results: Predictive accuracy comparison (robustness check)

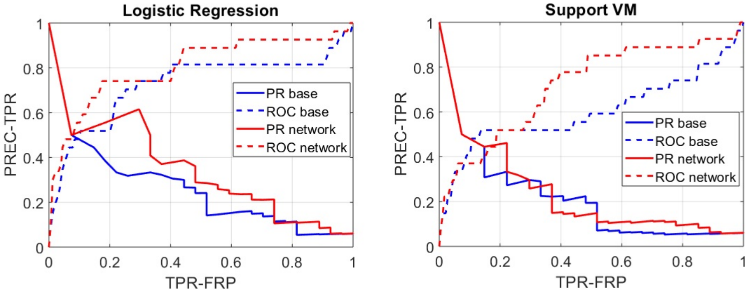


Figure: Precision Recall (PR) and Receiver Operating Characteristic (ROC) curves for the baseline credit risk models and for the network-augmented models. In each panel, dotted lines represent the ROC curves while solid lines refer to PR curves. In blue, we show the results related to the baseline models 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;
- This paper, data set used as well as the code to replicate the use case is available on our **fintech-ho2020 platform**.

Use Case II: Spatial regression models to improve P2P credit risk management

Arianna Agosto, Paolo Giudici and Thomas Leach

University of Pavia

Motivation

- Differently from banks, P2P platforms have limited access to historical data typically used in credit risk assessment.
- P2P platforms generate "alternative" data, which consist of the direct (disintermediated) transactions between their customer borrowers and/or lenders. These data can be used to assess credit risk and investigate contagion effects.
- Transactional data can also be used by banks, especially when evaluating new customers, or to assess contagion effects in networks of borrowers.

Background

- Financial network models (see, eg., Billio et al., 2012, Diebold and Yilmaz, 2014, Hautsch et al., 2015, Giudici and Spelta 2016) allow to study contagion effects but are often merely descriptive.
- Spatial econometric models (LeSage and Pace, 2009) can incorporate dependence among observations that are in a kind of proximity, not necessarily geographical.
- When applied to credit risk, spatial econometric models can be used both as a credit scoring model, to estimate the default risk of a given company, and as a contagion model (Calabrese et al., 2017).

Contribution

In this paper:

- We "simulate" transactions between the borrowers of a P2P lending platform using trade flows data between corporate sectors.
- Based on the proxied transactions, we apply a binary spatial model to measure credit risk of SMEs, taking contagion effects into account.

The model - I

Let Y a vector of binary dependent variables (default or not) and Y^* a vector of continuous underlying latent variables.
We consider the observation mechanism as

$$y_i = \begin{cases} 1, & \text{if } y_i^* > 0 \\ 0, & \text{otherwise} \end{cases}$$

with $i = 1, 2, \dots, n$

while the latent (default risk) mechanism has a spatial autoregressive structure:

$$Y^* = \rho W Y^* + X\beta + \epsilon$$

with

- X a $n \times k$ matrix of explanatory variables
- W the spatial lag weight matrix with ρ the associated coefficient
- ϵ the error term.

The model - II

The model implies heteroskedasticity in the errors:

$$Y^* = (I - \rho W)^{-1} X\beta + e$$

where

$$e = (I - \rho W)^{-1}\epsilon$$

thus

$$var(e) = \sigma_\epsilon^2 \left[(I - \rho W)'(I - \rho W)\right]^{-1}$$

Given the connectivity matrix W, the higher the ρ parameter

- the higher the spatial dependence between observations
- the higher the shocks covariance.

It can thus be interpreted as a *contagion* parameter.

Building the connectivity matrix - I

The World Input Output Trade (WIOT) statistics provide information on the aggregate trade volumes of 52 economic sectors in each country with all sectors in all countries.

For a given country, let A the sector of company i, B the sector of company j and f_{AB} the trade flow from sector A to sector B.

Replacing the individual flows with the aggregate (sectorial) ones, we obtain the approximate trade matrix F with entries:

$$f_{ij} = f_{AB} = \sum_{l \in A} \sum_{m \in B} f_{lm}$$

Building the connectivity matrix - II

To proxy the individual companies' flows, we first calculate the ratio between company i turnover $\tilde{x}_i$ and the sector A total turnover:

$$x_i = \frac{\tilde{x}_i}{\sum_{l \in A} \tilde{x}_l}$$

Then we calculate the ratio between company j turnover $\tilde{y}_i$ and the sector B total turnover:

$$y_j = \frac{\tilde{y}_j}{\sum_{m \in B} \tilde{y}_m}$$

The product $x_i y_j$ is a proxy of the proportion of flows from company i to company j on the total flows from sector A to sector B.

Building the connectivity matrix - III

Repeating this calculation for all companies and then computing the entrywise product with the trade matrix F, we get the connectivity matrix:

$$W = \begin{pmatrix} x_1 y_1 F_{1,1} & x_1 y_2 F_{1,2} & \cdots & x_1 y_n F_{1,n} \\ x_2 y_1 F_{2,1} & x_2 y_2 F_{2,2} & \cdots & x_2 y_n F_{2,n} \\ \vdots & \ddots & \dots & \vdots \\ x_n y_1 F_{n,1} & x_n y_2 F_{n,2} & \cdots & x_n y_n F_{n,n} \end{pmatrix}$$

The *ij* element can be interpreted as the proxy of the trade flow from company i to company j.

Conversely, the *ji* element can be interpreted as the proxy of the trade flow from company j to company i.

Application - Data

- Data collected from modeFinance, a European Credit Assessment Institution which supplies credit scoring to P2P platforms specialized in business lending.
- The complete dataset includes $\approx 15,000$ Italian SMEs, for which it is provided:
    - a set of financial ratios relative to accounting year 2015
    - information about the status of the company (0=Active, 1=Default) in 2016.
- From the available financial ratios, we select:
    - the return on equity ratio;
    - the activity ratio, expressed as the ratio between sales and total assets;
    - the solvency ratio, calculated as the ratio between the net income and the total debt.

Results

	$n = 1185$			
	Logit model		Binary SAR model	
	Estimate	Std. error	Estimate	Std. error
ρ	-	-	0.78	(0.23)
Constant	-2.11	(0.16)	0.44	(0.46)
Return On Equity	-0.69	(0.10)	-0.53	(0.15)
Activity Ratio	0.02	(0.10)	0.05	(0.13)
Solvency Ratio	-0.01	(0.00)	-0.03	(0.01)
AUC	0.798		0.806	

Table: Results of logit and binary SAR model estimation.

Results

	$n = 1185$		$n = 2000$		$n = 3000$	
	Estimate	Std. error	Estimate	Std. error	Estimate	Std. error
ρ	0.76	(0.16)	0.57	(0.17)	0.74	(0.16)
Constant	1.75	(0.93)	1.32	(0.75)	1.21	(0.55)
Total Assets/Total Liabilities	-1.06	(0.65)	-1.58	(0.50)	-1.11	(0.41)
(Current Assets-Stocks)/Current Liabilities	-0.89	(0.25)	-0.32	(0.20)	-0.31	(0.16)
EBITDA/Operating revenues	-1.87	(0.60)	-2.06	(0.47)	-2.84	(0.47)
Loss dummy for the period	1.42	(0.22)	1.43	(0.17)	1.36	(0.15)
AUC Binary SAR	0.832		0.828		0.825	
AUC Logit	0.827		0.827		0.824	

Table: Results of binary SAR model estimation.

Conclusions

- Simulating direct transactions between companies in a P2P platform, we have estimated a credit scoring model that includes a systemic component, based on transactions, and an idiosyncratic component, based on financial ratios.
- Our empirical findings show that the contagion parameter ρ is significant and > 0.5, for different sample sizes and sets of regressors.
- The improvement in model accuracy, even with respect to a well performing baseline specification, is positive.

Further research

- Sparse covariance structure: it is simplistic to assume that each company has trading relationships with all the others.
- Proximity measure: the distance may be calculated in terms of net flows.
- Does the contagion parameter change in time and under stress conditions?