This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 825215. All material presented here reflects only the authors' view. The European Commission is not responsible for any use that may be made of the information it contains.

Big data analytics Use Cases

FIN-TECH HO2020 project

February 2, 2020

Table of contents

Use case I: Network based scoring models to improve credit risk management in P2P lending (Giudici, Hadji-Misheva and Spelta, UNIPV and ZHAW, 2019)

Peer to peer lending

- Among FinTech applications that rely on big data analytics, innovative ones are those based on peer-to-peer (P2P) financial transactions, such as peer to peer lending, crowdfunding and invoice trading.
- The concept peer-to-peer captures the interaction between individual units, which eliminates the need for a central intermediary.

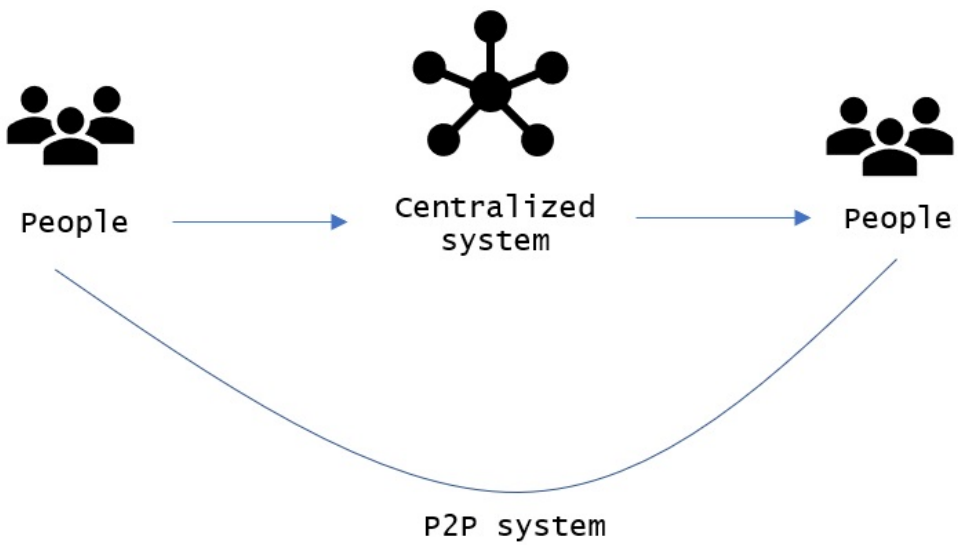


Figure 1: P2P Lending Platforms

Peer to Peer Lending: Opportunities

- improved rates of return compared to those available on bank deposits, together with relatively low fees for borrowers;
- improved financial inclusion - P2P platforms are able to provide financing to borrowers unable to access bank lending;
- improved quality, speed of service and user experience to both borrowers and lenders.

Peer to Peer Lending: Risks

- While both classic banks and P2P platforms rely on credit scoring models for the purpose of estimating the credit risk of their loans, the incentive for model accuracy may differ significantly:
  - In a bank, the assessment of credit risk of the loans is conducted by the financial institution itself which, being the actual entity that assumes the risk, is interested to have the most accurate possible model.
  - In a P2P lending platform, credit risk of the loans is determined by the platform but the risk is fully borne by the lender.
- Another factor that penalizes the accuracy of P2P credit scoring models is that they are (still) based on limited data.

Peer to peer lending: Proposal

- P2P platforms operate as social networks, which involve their users and, in particular, the borrowers, in a continuous networking activity.
- From the P2P platform perspective, network data models should be employed, to improve credit risk measurement accuracy.
- From a supervisory viewpoint, the modelling of network data should be allowed, so that P2P lenders produce more reliable credit risk estimates.

Peer to peer lending: use cases data

- The proposed models are built and tested on data that concerns 15,045 European SMEs (potential costumers of P2P platforms), for which we observe 24 financial indicators from the balance sheets of the year 2015 , and the corresponding one-year lagged 2016 status variable.
- The response variable indicates the legal status of a company: [0=active and 1=default]

Status	Frequency	Percentage
Active [0]	13'413	89.20%
Default [1]	1'632	10.80%

Table 1: Distribution of the Response Variable

Data - predictors

Var	Formula (Description)	Active Mean	Defaulted Mean
V_1	(Total Assets - Shareholders Funds)/Shareholders Funds	8.87	9.08
V_2	(Longterm debt + Loans)/Shareholders Funds	1.25	1.32
V_3	Total Assets/Total Liabilities	1.51	1.07
V_4	Current Assets/Current Liabilities	1.6	1.06
V_5	(Current Assets - Current assets: stocks)/Current Liabilities	1.24	0.79
V_6	(Shareholders Funds + Non current liabilities)/Fixed Assets	8.07	5.99
V_7	EBIT/Interest paid	26.39	-2.75
V_8	(Profit (loss) before tax + Interest paid)/Total Assets	0.05	-0.13
V_9	P/L after tax/Shareholders Funds	0.02	-0.73
V_{10}	Operating Revenues/Total Assets	1.38	1.27
V_{11}	Sales/Total Assets	1.34	1.25
V_{12}	Interest Paid/(Profit before taxes + Interest Paid)	0.21	0.08
V_{13}	EBITDA/Interest Paid	40.91	5.71
V_{14}	EBITDA/Operating Revenues	0.08	-0.12
V_{15}	EBITDA/Sales	0.09	-0.12
V_{16}	Constraint EBIT	0.13	0.56
V_{17}	Constraint PL before tax	0.16	0.61
V_{18}	Constraint Financial PL	0.93	0.98
V_{19}	Constraint P/L for period	0.19	0.64
V_{20}	Trade Payables/Operating Revenues	100.3	139.30
V_{21}	Trade Receivables/Operating Revenues	67.59	147.12
V_{22}	Inventories/Operating Revenues	90.99	134.93
V_{23}	Total Revenue	3557	2083
V_{24}	Industry Classification on NACE code	4566	4624
	Total number of institutions (%)	13413 (89.15%)	1632 (10.85%)

Table 2: Summary Statistics of all variables, split by status

Proposal

- We investigate whether network information can improve loan default predictions and further protect lenders, in a financial stability context.
- To achieve this aim we build a network model from the available platform networking data. This is available in a multi-layer perspective, e.g.:
  - financial transactions;
  - economic similarities;
  - trade flows.

Scoring Models

- Logistic regression is the most widely used method to estimate default probabilities:

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

  where p_i 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 is:

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

Distance networks

- Consider all predictor variables and (without loss of generality) a sample of 4514 SMEs (default rate 11%).
- Given two borrower companies whose financial characteristics are collected in vectors $\mathbf{y}_i$ and $\mathbf{y}_j$, at any time point their distance $d_{i,j}$ can be calculated as:

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

  where Δ is a diagonal matrix whose i-th diagonal element represents the standard deviation of the series.
- The distances can be embedded into a $N \times N$ dissimilarity matrix $\mathbf{D}$.

Minimal Spanning Tree Networks

- To simplify the network, likely to be fully connected when so many variables are being considered, we employ the Minimal Spanning Tree (MST), a tree of $N-1$ edges that is obtained from hierarchical clustering of the nodes:
- Initially N clusters that corresponds to the N borrower companies are considered. Then, at each step, two clusters l_i and l_j are merged into a single cluster if:

$$d(l_i, l_j) = \min\{d(l_i, l_j)\}$$

with the distance between clusters being defined as:

$$d(l_i, l_j) = \min\{d_{pq}\}$$

with $p \in l_i$ and $q \in l_j$.

Network based logistic regression

- Based on the thresholded MST representation we calculate centrality measures, g_{ik} .
- We then build a network based logistic regression model.

$$ln(\frac{p_i}{1-p_i}) = \alpha_c + \sum_j \beta_j x_{ij} + \sum_k \gamma_k g_{ik},$$

- For robustness we also consider predictive models alternative to the logistic regression.

Results: Borrowers' network

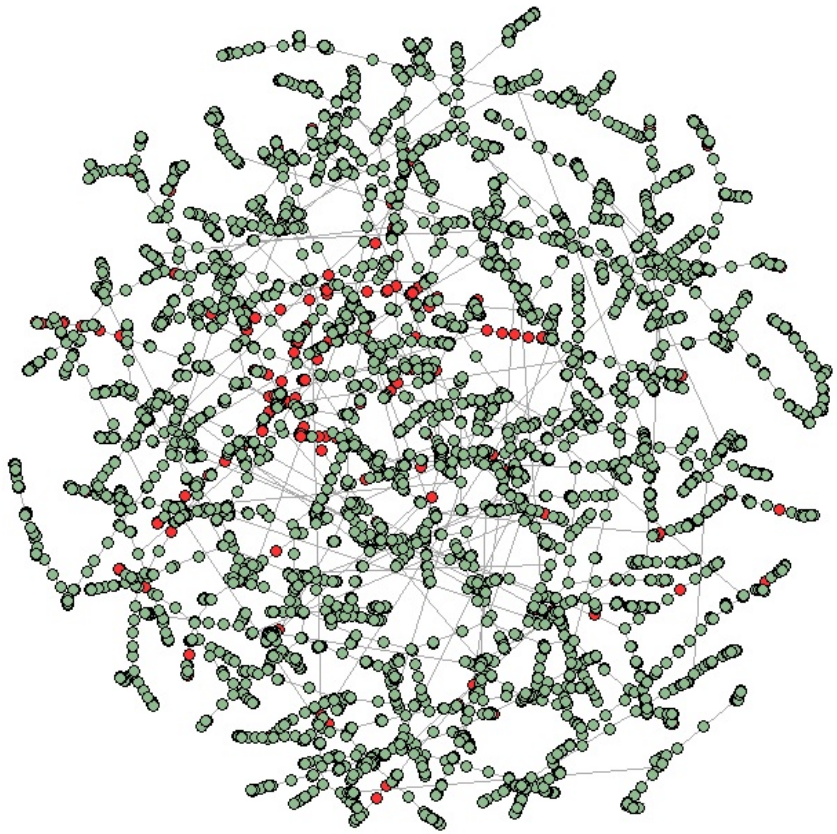


Figure 2: Minimal spanning tree of the borrowers' network.

Results: Predictive performance

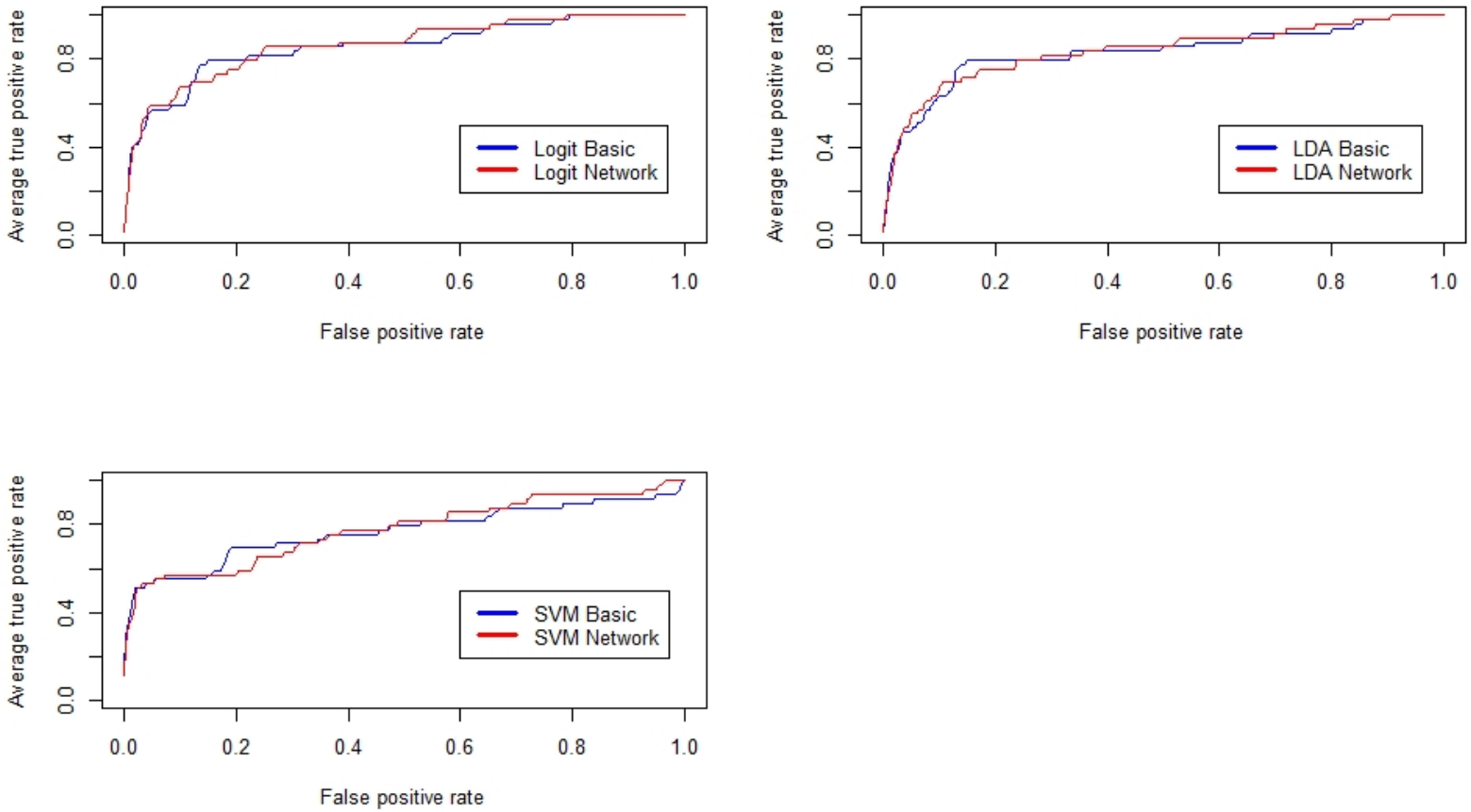


Figure 3: Receiver Operating Characteristic (ROC) curves for the baseline credit risk models (blue) and for the network-augmented models (red). Network models further improve predictive accuracy: for the logistic regression model, AUROC increases from 0.79 (baseline) to 0.81 (network based)

Use case II: Factorial network models to improve P2P credit risk management (Ahelegbey, Giudici and Hadji-Misheva, UNIPV and ZHAW, 2019)

Proposal

- In this use-case we propose to improve credit risk for P2P platforms by means of network clustering.
- We assume that a latent factor model will partition companies in two groups: the interconnected ones, affected by contagion; and the disconnected ones, not affected by contagion.

Latent Factor Models

- Let X be the $n \times m$ observed data matrix. X can be expressed as a factor model given by

$$X = FW' + \varepsilon$$

where F is $n \times k$ matrix of latent factors, W is $p \times k$ matrix of factor loadings and ε is $n \times p$ matrix of errors.

- We then construct a network in which the probability of a link between nodes-i and j is given by

$$\gamma_{ij} = P(G_{ij} = 1|F) = \Phi[\theta + (FF')_{ij}]$$

where Φ is the standard normal cumulative density function and $\theta \in \mathbb{R}$ is a network density parameter.

Latent network models - I

- We place a link between nodes i and j when $(\gamma_{ij} > \gamma)$, with $\gamma \in (0,1)$ a threshold parameter.
- To broaden the robustness of the results, we compare different threshold values of $\gamma = (0.01, 0.05, 0.1)$.

Latent network models - II

- To make the model more parsimonious, we consider the Lasso approach, which minimises the penalized log-likelihood function:

$$\mathscr{L}_\lambda = \sum_{i=1}^{n}\sum_{j=1}^{p}\Big[Y_i(\beta_0 + X_{ij}\beta_j) - \log\left(1+\exp(\beta_0+X_{ij}\beta_j)\right)\Big] - \lambda\sum_{j=0}^{p}|\beta_j|$$

- with n the number of observations, p the number of predictors, and λ is the penalty term, such that large values of λ shrinks a large number of the coefficients towards zero.

Results -Choosing the latent space dimension

No.	Eigenvalue	Variance Explained (%)	Cumulative (%)
1	5.18	21.60	21.60
2	2.58	10.73	32.33
3	2.50	10.41	42.74
4	1.60	6.69	49.42
5	1.42	5.92	55.34
6	1.30	5.40	60.74
7	1.16	4.82	65.55
8	1.09	4.56	70.11
9	0.99	4.11	74.22
10	0.93	3.88	78.10
11	0.80	3.35	81.45
12	0.79	3.31	84.76
13	0.75	3.11	87.87
14	0.56	2.35	90.22
15	0.53	2.21	92.43
16	0.51	2.12	94.55
17	0.43	1.80	96.35
18	0.37	1.54	97.89
19	0.17	0.69	98.58
20	0.11	0.47	99.05
21	0.09	0.36	99.41
22	0.07	0.27	99.68
23	0.06	0.26	99.94
24	0.01	0.06	100.00

Table 3: The eigenvalues of the singular value decomposition to determine the factors to retain.

Results - All borrowers

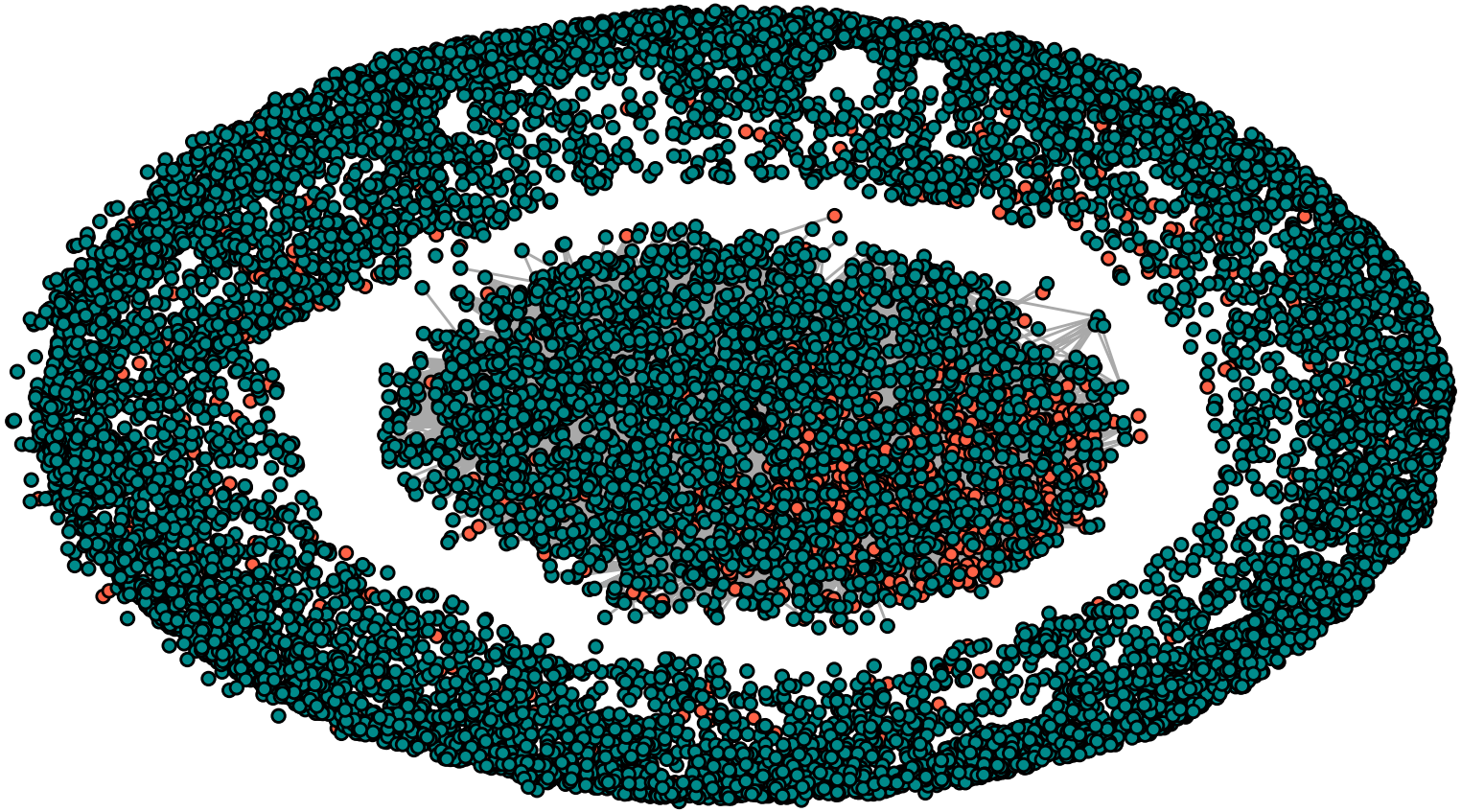


Figure 4: The estimated factor network structure of all borrowers (15045 companies, default rate = 10.80%)

Results - Connected borrowers

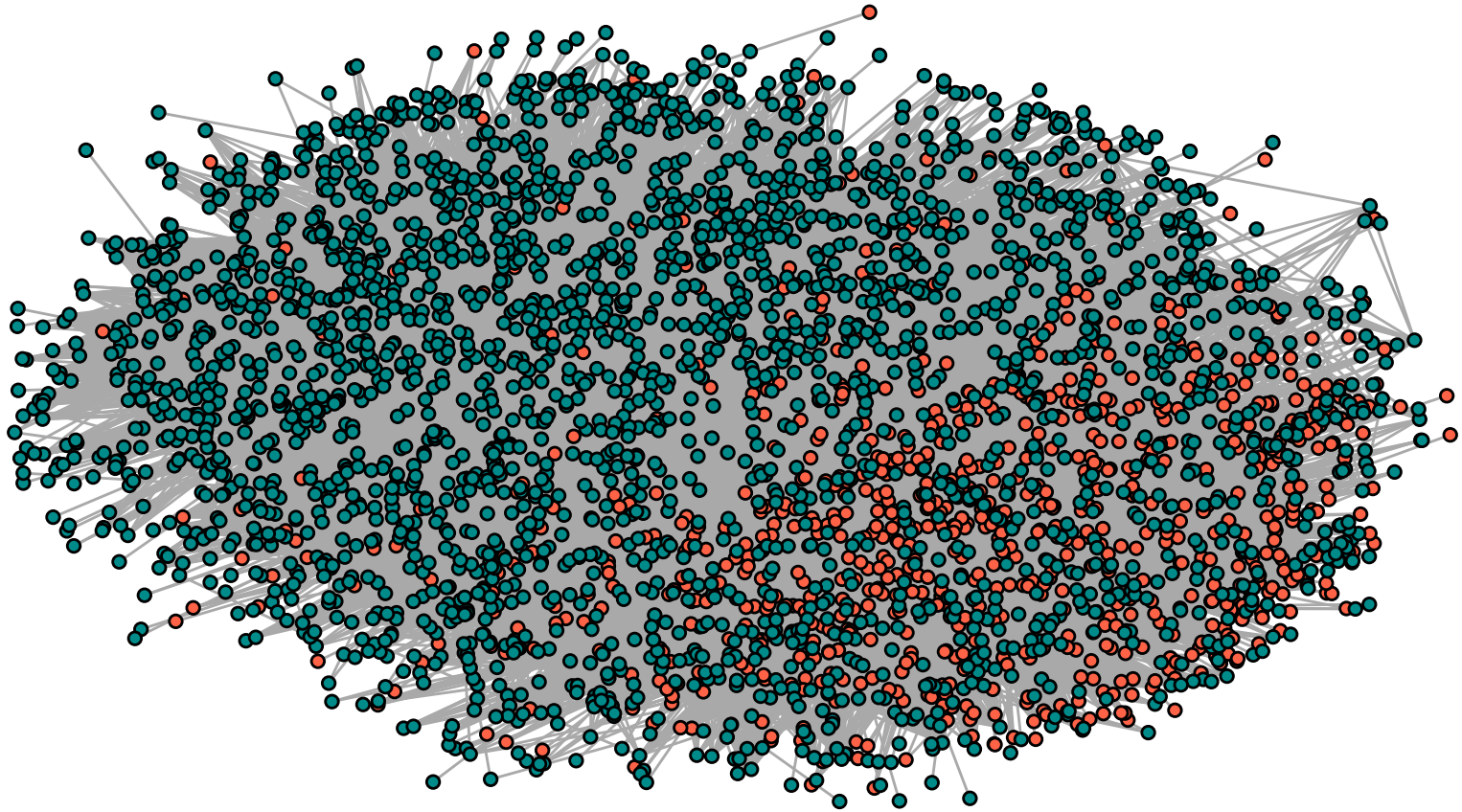


Figure 5: The estimated factor network structure for the connected borrowers. (4305 companies, default rate = 22.40%)

Results - Significant variables

	All	Connected	Not connected
(Intercept)	-1.961	-1.811	-1.126
V_1	0.003	0	0.002
V_2	0	0.009	0
V_3	-0.562	-0.348	-1.237
V_4	-0.298	-0.106	-0.437
V_5	0.003	0	0
V_6	0.003	0	0.005
V_7	0.004	0	0
V_8	-2.683	-2.322	0.519
V_9	-0.045	0.042	-0.576
V_{10}	-0.145	0.023	0
V_{11}	0.202	0	0.035
V_{12}	0.060	0.038	0.033
V_{13}	-0.003	0	0
V_{14}	-0.177	-0.400	0
V_{15}	-0.360	-0.174	0
V_{16}	0.155	0.726	0
V_{17}	0.538	0.412	0.398
V_{18}	0.167	0.256	0.025
V_{19}	0.594	0.065	0.492
V_{20}	0.0001	0	0.001
V_{21}	0.002	0.001	0.003
V_{22}	0.001	0	0.001
V_{23}	-0.00003	-0.00001	-0.00004
V_{24}	-0.00000	-0.00003	0.00001

Table 4: Comparing the Lasso logistic model estimated coefficients in the different clusters: i) all companies; ii) only connected companies; iii) only non-connected companies.

Predictive Modeling

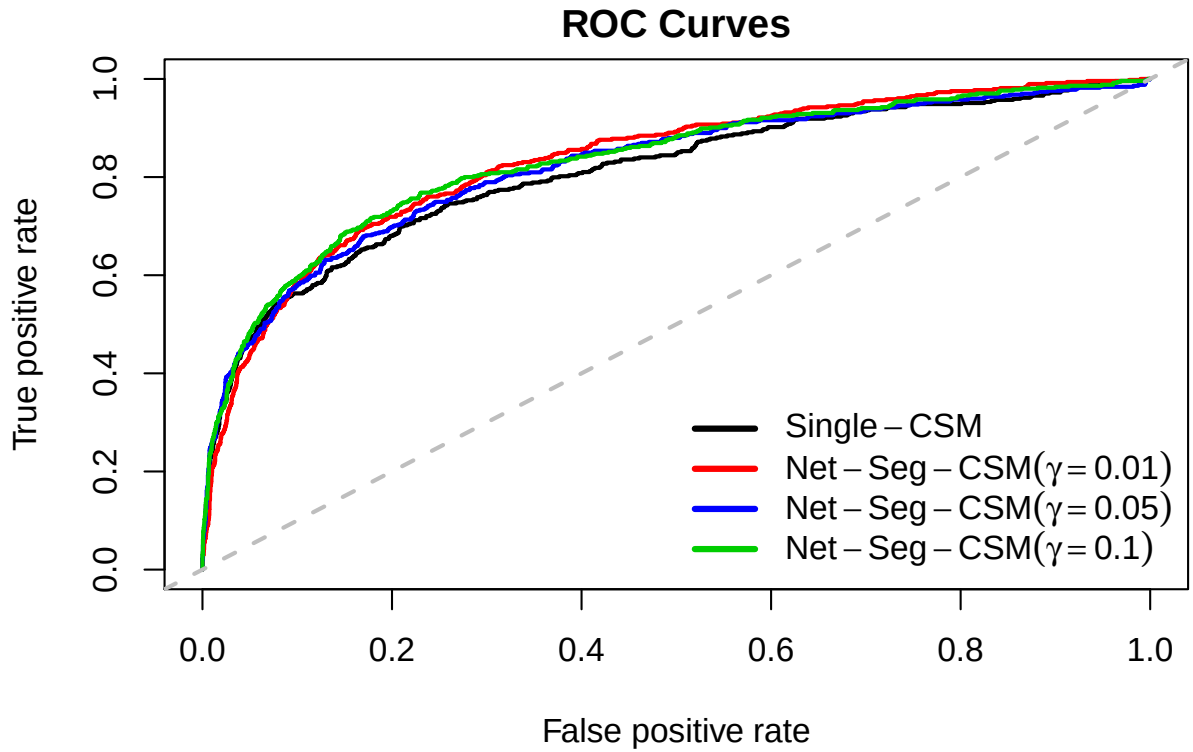


Figure 6: Plot of the ROC curves of the single credit score model (Single-CSM): AUROC = 0.80; and the network-based segmented credit score models (Net-Seg-CSM): AUROC= 0.82, for threshold values of $\gamma = \{0.01, 0.05, 0.1\}$.

Use case III: Spatial regression models to improve P2P credit risk management (Agosto, Giudici and Leach, UNIPV, 2019)

Proposal

- In this use case we employ rather than "statistical" network data a proxy to "physical" network data.
- We correspondingly employ spatial econometrics models to incorporate contagion, following Calabrese, Elkink and Giudici (2017).
- Spatial econometrics can incorporate dependence among observations that are in any kind of proximity, not only geographical.

Spatial Autoregressive models - I

- We assume the default process to be represented by a binary dependent variable, y, which depends on a continuous underlying latent variable y^*, so that $y = 1$ if $y^* > 0$ and 0 otherwise. Example: logit transformation in logistic regression models.
- We then assume:

$$Y^* = \rho W Y^* + X\beta + \varepsilon,$$

where Y^* is a continuous random vector, X is an $n \times k$ matrix of explanatory variables and W is the spatial weight matrix with ρ the associated contagion parameter.

Spatial Autoregressive models - II

- The model implies heteroskedastic errors **e** as follows:

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

where

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

and

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

- To estimate ρ and β we have developed a generalised method of moments algorithm, that uses instrumental variables. W is instead assumed to be known, and calculated from the Input-Output trade matrices.

Spatial weighting matrix - I

- The W matrix is based on an exogenously defined network, where the nodes correspond to companies and the links express the volume of trade between any pair of companies.
- This information is generally not available, and we approximate it using data on aggregate input-output trade between sectors from the World Input Output Trade database.
- For a given country, define A as the sector of company i, B as the sector of company j, and let f_{AB} be the trade flow from sector A to sector B, while f_{BA} is the trade flow from sector B to sector A.
- Replacing the individual flows with the aggregate ones, the entries of the approximate trade matrix F are then obtained as:

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

Spatial weighting matrix III

- The product $\bar{x}\bar{y}$ proxies the proportion of flows from company i to company j on the total flows from A to B:

$$R = \langle \bar{x}, \bar{y} \rangle = \begin{pmatrix} \bar{x}_1\bar{y}_1 & \bar{x}_1\bar{y}_2 & \cdots & \bar{x}_1\bar{y}_n \\ \bar{x}_2\bar{y}_1 & \bar{x}_2\bar{y}_2 & \cdots & \bar{x}_2\bar{y}_n \\ \vdots & \ddots & \cdots & \vdots \\ \bar{x}_n\bar{y}_1 & \bar{x}_n\bar{y}_2 & \cdots & \bar{x}_n\bar{y}_n \end{pmatrix}$$

- Calculating the entrywise product of R with the trade matrix F, we obtain the weight matrix

$$W = R \circ F = \begin{pmatrix} \bar{x}_1\bar{y}_1 R_{1,1} & \bar{x}_1\bar{y}_2 R_{1,2} & \cdots & \bar{x}_1\bar{y}_n R_{1,n} \\ \bar{x}_2\bar{y}_1 R_{1,1} & \bar{x}_2\bar{y}_2 R_{2,2} & \cdots & \bar{x}_2\bar{y}_n R_{2,n} \\ \vdots & \ddots & \cdots & \vdots \\ \bar{x}_n\bar{y}_1 R_{n,1} & \bar{x}_n\bar{y}_2 R_{n,2} & \cdots & \bar{x}_n\bar{y}_n R_{n,n} \end{pmatrix}$$

in which the ij element is the proportion of trade flow from i to j and the ji element the proportion of trade flows from j to i.

Estimation - I

Homoskedastic errors are a restrictive assumption. From the reduced form of the equation we have:

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

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

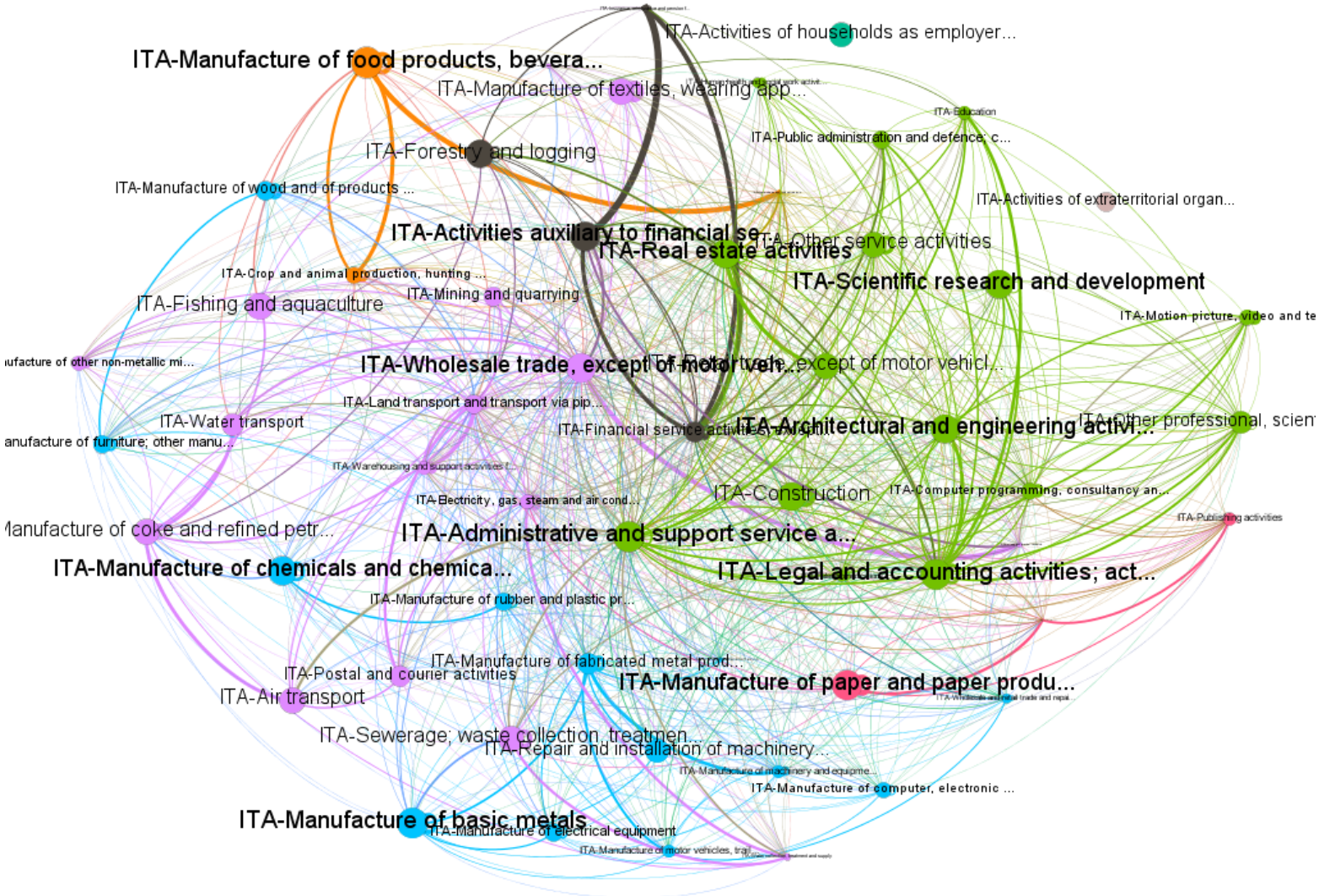
$$Var(e) \propto \left[(I - \rho W)'(I - \rho W) \right]^{-1}$$

The spatial lag parameter, ρ, reflects the spatial dependence inherent in the sample data, measuring the average influence of neighbouring observations on observations in vector Y.

Data

- We extract from the available database a sample of 1185 peer to peer lending borrowing companies, operating in Italy. The percentage of defaulted companies is close to 12%
- Correspondingly, we download World Input Output Trade Data for the 52 sectors of Italy, in the year 2015.

Results - Trade network between sectors



Results - Estimation I

We select three financial ratios, reflecting: operational performance, business sustainability and financial sustainability. Specifically, we consider:

- the return on equity ratio (RATIO012)
- the activity ratio, expressed as the ratio between sales and total assets (RATIO018);
- the solvency ratio, expressed as the ratio between the net income and the total debt (RATIO027)

Baseline logistic regression:

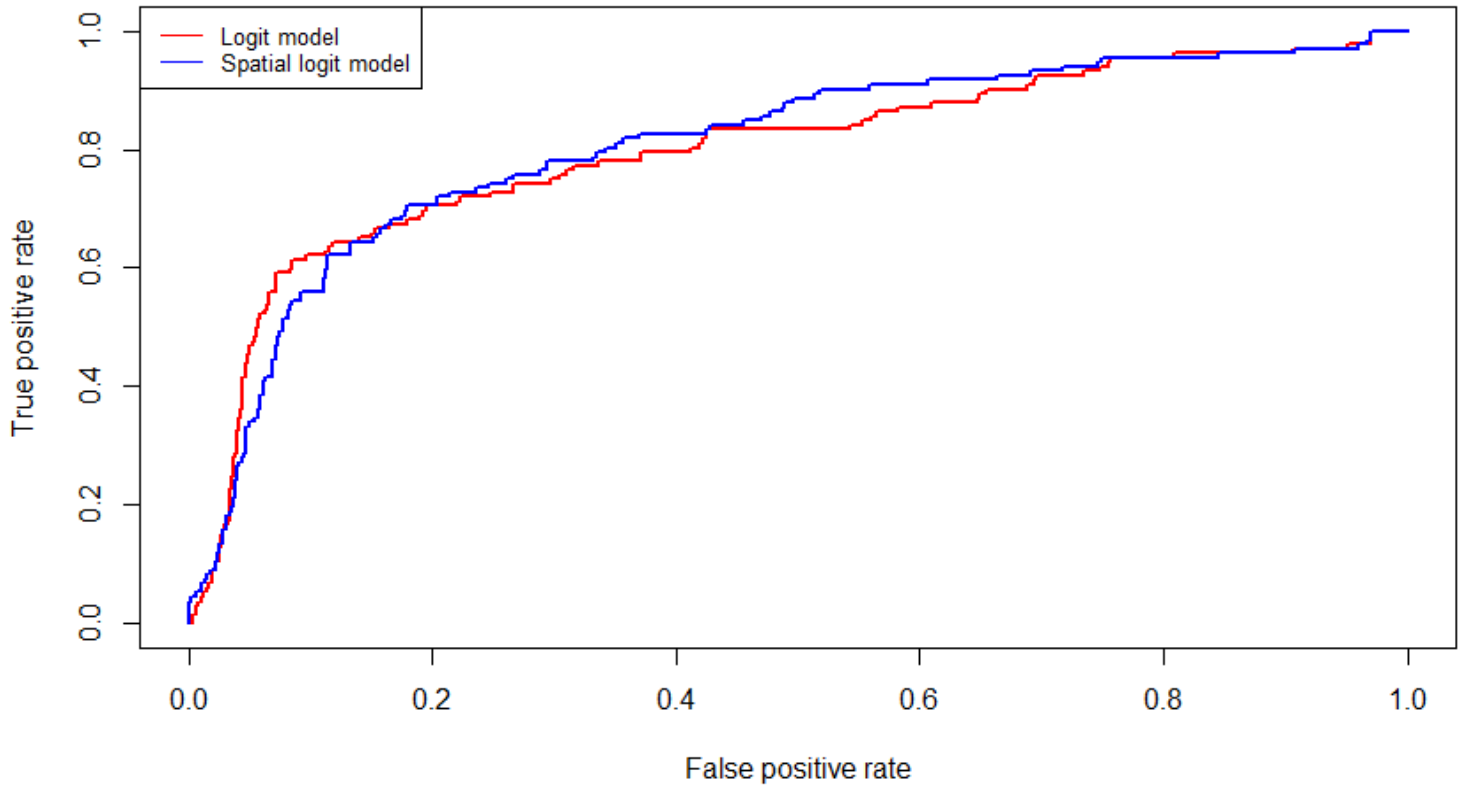
	Estimate	Std. Error	$Pr(>\|z\|)$
Intercept	-2.11	0.16	2.97e-38
β_1 (RATIO012)	-0.69	0.10	6.35e-11
β_2 (RATIO018)	0.02	0.10	0.84
β_3 (RATIO027)	-0.01	0.00	9.10e-04

Results - Estimation II

Spatial logistic model:

	Estimate	Std. Error	Pr($>\|z\|$)
ρ	0.78	0.23	5.44e-04
Intercept	0.44	0.46	0.35
β_1 (RATIO012)	-0.53	0.15	2.24e-04
β_2 (RATIO018)	0.05	0.13	0.69
β_3 (RATIO027)	-0.03	0.01	0.03

Results - ROC



Receiver Operating Characteristic (ROC) curves for the baseline credit risk models (red) and for the spatialli-augmented models (blue). Network models further improve predictive accuracy: for the logistic regression model, AUROC increases from 0.79 (baseline) to 0.81 (spatial based)