Twitter data models for bank risk contagion

Paola Cerchiello, Paolo Giudici, Giancarlo Nicola

Abstract

A very important and timely area of research in finance is systemic risk modelling, which concerns the estimation of the relationships between different financial institutions, with the aim of establishing which of them are more contagious/subject to contagion. The aim of this paper is to develop a systemic risk model which, differently from existing ones, employs not only the information contained in financial market prices, but also big data coming from financial tweets. From a methodological viewpoint, we propose a new framework, based on graphical Gaussian models, that can estimate systemic risks with stochastic network models based on two different sources: financial markets and financial tweets, and suggest a way to combine them, using a Bayesian approach. From an applied viewpoint, we present the first systemic risk model based on big data, and show that such a model can help predicting the default probability of a bank, conditionally on the others.

1. Introduction

Systemic risk models address the issue of interdependence between financial institutions and, specifically, measure how bank default risks are transmitted among banks.

The study of bank defaults is important for two reasons. First, an understanding of the factors related to bank failure enables regulatory authorities to supervise banks more efficiently. If supervisors can detect problems early enough, regulatory actions can be taken, to prevent a bank from failing and, therefore, to reduce the costs of its bail-in, faced by shareholders, bondholders and depositors; or those of its bail-out, faced by governments and, ultimately, by the taxpayers. Second, the failure of a bank very likely induces failures of other banks or of parts of the financial system. Understanding the determinants of a single bank failure may thus help to understand the determinants of financial systemic risks, were they due to microeconomic idiosyncratic factors or to macroeconomic imbalances. When problems are detected, their causes can be removed or isolated, to limit "contagion effects".

Most research papers on bank failures are based on financial market models, that originate from the seminal paper of Merton (1974), in which the market value of bank assets is matched against bank liabilities. Due to its practical limitations, Merton's model has been evolved into a reduced form (see e.g. Vasicek, 1984), leading to a widespread diffusion of the resulting approach, and the related implementation in regulatory models.

The last few years have witnessed an increasing research literature on systemic risk, with the aim of identifying the most contagious institutions and their transmission channels. Specific measures of systemic risk have been proposed for the banking sector; in particular, by Acharya et al. (2010), Adrian and Brunnermeier (2011), Brownlees and Engle (2012), Acharya et al. (2012), Dumitrescu and Banulescu (2014) and Hautsch et al. (2015). On the basis of market prices, these authors calculate the quantiles of the estimated loss probability distribution of a bank, conditional on the occurrence of an extreme event in the financial market.

The above approach is useful to establish policy thresholds aimed, in particular, at identifying the most systemic institutions. However, it is a bivariate approach, which allows to calculate the risk of an institution conditional on another (or on a reference market), but it does not address the issue of how risks are transmitted between different institutions in a multivariate framework.

Trying to address the multivariate nature of systemic risk, researchers have proposed a network modelling approach, following the idea in Diamond and Dybvig (1983) and the seminal papers of Sheldon and Maurer (1998), Eisenberg and Noe (2001), Boss et al. (2004), Upper and Worms (2004). In this literature, interconnectedness is related to the detection of the most central players in a network that describes financial flows between agents. While the simplest way of measuring

the centrality of a node in the network is by counting the number of neighbors that it has, more sophisticate measures of centrality have been applied, including that shown in Battiston et al. (2012) who develop a network algorithm -the DebtRank- starting from Google's PageRank algorithm.

A different type of network models, recently proposed, are based on correlations (or distances) between financial descriptors of agents, such as their stock market prices, bond interest rate spreads or corporate default spreads. The first contributions in this framework are Mantegna (1999), Onnela et al. (2004), Tumminello et al. (2004) and, recently, Billio et al. (2012) and Diebold and Yilmaz (2014), who propose measures of connectedness based on Granger-causality tests and variance decompositions. Barigozzi and Brownlees (2013), Ahelegbey et al. (2015) and Giudici and Spelta (2016) have extended the approach introducing stochastic graphical models.

Here we shall follow this latter approach, and add a stochastic framework, based on graphical models. We will thus be able to derive, on the basis of market price data on a number of financial institutions, the network model that best describes their interrelationships and, therefore, explains how systemic risk is transmitted among them.

It is well known that market prices are formed in complex interaction mechanisms that often reflect speculative behaviours, rather than the fundamentals of the companies to which they refer. Market models and, specifically, financial network models based on market data may, therefore, reflect "spurious" components that could bias systemic risk estimation. This weakness of the market suggests to enrich financial market data with data coming from other, complementary, sources. Indeed, market prices are only one of the evaluations that are carried out on financial institutions: other relevant ones include ratings issued by rating agencies, reports of qualified financial analysts, and opinions of influential media.

Most of the previous sources are private, not available for data analysis. However, summary reports from them are now typically reported, almost in real time, in social networks and, in particular, in tweets. In parallel with these developments, seminal papers on the statistical analysis of such data have recently appeared: see, for example, Bollen et al. (2011), Bordino et al. (2012), Choi et al. (2012), Feldman (2013), Cerchiello and Giudici (2015), Andersen (2016)), who all show the added value of tweets and, more generally, of textual data, in economics and finance.

Indeed twitter data offers the opportunity to extract data that can complement market prices and that can, in addition, "replace" market information when not available (as it occurs for banks that are not listed).

To extract from tweets data that can be assimilated to market prices, their text has to be preprocessed using semantic analysis techniques. In our context, if financial tweets on a number of banks are collected daily, semantic analysis allows to obtain a daily "sentiment" that expresses, for each day, how each considered bank is, on average, being evaluated by twitterers.

In this paper we propose to build graphical Gaussian models using daily variation of bank "sentiment", and to integrate them with graphical models based on market data, by means of a Bayesian approach. This allows to obtain a comprehensive measurement framework of bank interconnectedness, that can be employed to understand contagion effects.

The novelty of this paper is twofold. From a methodological viewpoint, we propose a framework, based on graphical Gaussian models, that can estimate systemic risks with models based on two different sources: financial markets and financial tweets, and suggest a way to combine them, using a Bayesian approach.

From an applied viewpoint, we propose a novel usage of big data contained in financial tweets, and show that such data can shed further light on the interrelationships between financial institutions.

The rest of the paper is organised as follows: in Section 2 we introduce our proposal; in Section 3 we apply our proposal to financial and tweet data on the Italian banking market and, finally, in Section 4 we present some concluding remarks.

2. Methodology

We first introduce the graphical network models that will be used to estimate relationships between banks, both with market and tweet data.

Relationships between banks can be measured by their partial correlation, that expresses the direct influence of a bank on another. Partial correlations can be estimated assuming that the observations follow a graphical Gaussian model, in which Σ is constrained by the conditional independences described by a graph (see e.g. Lauritzen, 1996).

More formally, let $X = (X_1, ..., X_N) \in R^N$ be a $N-$dimensional random vector distributed according to a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$. Without loss of generality, we will assume that the data are generated by a stationary process, and, therefore, $\mu = 0$. In addition, we will assume throughout that the covariance matrix Σ is not singular.

Let $G = (V, E)$ be an undirected graph, with vertex set $V = \{1, ..., N\}$, and edge set $E = V \times V$, a binary matrix, with elements e_{ij}, that describe whether pairs of vertices are (symmetrically) linked between each other ($e_{ij} = 1$), or not ($e_{ij} = 0$). If the vertices V of this graph are put in correspondence with the random variables $X_1, ..., X_N$, the edge set E induces conditional independence on X via the so-called Markov properties (see e.g. Lauritzen, 1996).

In particular, the pairwise Markov property determined by G states that, for all $1 \leq i < j \leq N$:

$$e_{ij} = 0 \iff X_i \perp X_j | X_{V \setminus \{i,j\}}; \tag{1}$$

that is, the absence of an edge between vertices i and j is equivalent to independence between the random variables X_i and X_j, conditionally on all other variables $x_{V \setminus \{i,j\}}$.

Let the elements of Σ^{-1}, the inverse of the variance-covariance matrix, be indicated as $\{\sigma^{ij}\}$, Whittaker (1990) proved that the following equivalence also holds:

$$X_i \perp X_j | X_{V \setminus \{i,j\}} \iff \rho_{ijV} = 0 \tag{2}$$

where

$$\rho_{ijV} = \frac{-\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}} \tag{3}$$

denotes the ij-th partial correlation, that is, the correlation between X_i and X_j, conditionally on the remaining variables $X_{V \setminus \{i,j\}}$.

Therefore, by means of the pairwise Markov property, and given an undirected graph $G = (V, E)$, a graphical Gaussian model can be defined as the family of all N-variate normal distributions that satisfies the constraints induced by the graph on the partial correlations, as follows:

$$e_{ij} = 0 \iff \rho_{ijV} = 0 \tag{4}$$

for all $1 \leq i < j \leq N$.

Stochastic inference in graphical models may lead to two different types of learning: structural learning, which implies the estimation of the graphical structure G that best describes the data, and quantitative learning, that aims at estimating the parameters of a graphical model, for a given graph.

Structural learning can be achieved choosing the graphical structure with maximal likelihood. To this aim, we now recall the expression of the likelihood of a graphical Gaussian model.

For a given graph G, consider a sample X of size n. For a subset of vertices $A \subset N$, let Σ_A denote the variance-covariance matrix of the variables in X_A, and define with S_A the corresponding observed variance-covariance sub-matrix.

When the graph G is decomposable (and we will assume so) the likelihood of the data, under a graphical Gaussian model, nicely decomposes as follows (see e.g. Dawid and Lauritzen, 1993):

$$p(X|\Sigma, G) = \frac{\prod_{C \in \mathcal{C}} p(X_C|\Sigma_C)}{\prod_{S \in \mathcal{S}} p(X_S|\Sigma_S)}, \tag{5}$$

where X_C and X_S respectively denote the set of random variables belonging to the cliques and to the separators of the graph G, and where:

$$P(X_C|\Sigma_C) \propto |\Sigma_C|^{-n/2} exp[-\frac{1}{2} tr\left(S_C \, (\Sigma_C)^{-1}\right) \tag{6}$$

and similarly for $P(X_S|\Sigma_S)$.

Operationally, a model selection procedure compares different G structures by calculating the previous likelihood substituting for Σ its maximum likelihood estimator under G. For a complete (fully connected) graphical Gaussian model such an estimator is simply the observed variance-covariance matrix. For a general (decomposable) incomplete graph, an iterative procedure, based on the clique and separators of a graph, must be undertaken (see e.g. Lauritzen, 1996).

Through model selection, we obtain a graphical model that can be used to describe relationships between banks and, specifically, to understand how risks propagate in a systemic risk perspective.

Cerchiello and Giudici (2015) and Giudici and Spelta (2015) have shown, respectively in the context of country financial flows and bank returns, that Graphical Gaussian models are well suited to estimate interconnections between a large set of financial institutions, on the basis, respectively, of the available inter-country bank liability data or financial market data.

In our context, we have the additional task of selecting a graphical model for two different data sources: not only market data on banks but also big data, coming from financial tweets on the same banks. Indeed, the two data sources should be combined into a single one, before performing model selection. This is the additional contribution of the present paper, and can be achieved within a Bayesian framework, characterised by an Empirical Bayes appraoch to the specification of the prior distribution.

Empirical Bayes models (see e.g. Casella and George, 1985 and Carlin et al., 2000) adress the issue of specifying the prior distribution, an often controversial subject in Bayesian modelling, not on a priori ground,

but using data, assumed to come from a population different from the one considered as the main object of the statistical inference. In our context, the main object of inference is the correlation structure of market prices, which can be summarised in the correlation matrix parameter. Eliciting a prior distribution on a correlation matrix is a rather complex task, especially when a large number of variables is involved. Furthermore, even when feasible, such a prior may be highly influential on final inferences, possibly distorting Bayesian estimates toward the prior, rather than towards the actual data (see e.g. Casella and George,1985, Carlin et al., 2000; and, in a financial context, Giudici, 2001). The Empirical Bayes approach offers a possible solution to this problem, allowing the prior distribution to be also estimated from real data, possibly different from what used as main object of the inference. In our context, such data is available from Twitter and, therefore, it can be employed to estimate an "a priori" correlation matrix, based on sentiment data, to be combined, in a Bayesian model, with the market price correlation matrix.

More formally, we first specify a prior distribution for the parameter Σ. Dawid and Lauritzen (1993) propose a convenient prior, the hyper inverse Wishart distribution.

The hyper inverse Wishart distribution can be obtained from a collection of clique specific marginal inverse Wishart as follows:

$$l(\Sigma) = \frac{\prod_{C \in \mathcal{C}} l(\Sigma_C)}{\prod_{S \in \mathcal{S}} l(\Sigma_S)}, \tag{7}$$

where $l(\Sigma_C)$ is the density of an inverse Wishart distribution:

$$l(\Sigma_C) = \frac{|T_C|^{\frac{\alpha}{2}}}{2^{\frac{\alpha p}{2}} \Gamma_p(\frac{\alpha}{2})} |\Sigma_C|^{-\frac{\alpha+p-1}{2}} \exp(-1/2) tr(T_C \Sigma_C^{-1}) \tag{8}$$

with hyperparameters T_C and α, and similarly for $l(\Sigma_S)$. For the definition of the hyperparameters here we follow Giudici and Green (1999) and let T_C and T_S be the submatrices of a larger "scale" matrix T_0 of dimension $N \times N$, and choose $\alpha > N$.

Dawid and Lauritzen (1996) and Giudici and Green (1999) show that, under the previous assumptions, the posterior distribution of the variance-covariance matrix Σ is a hyper Wishart distribution with $\alpha + n$ degrees of freedom and a scale matrix given by:

$$T_n = T_0 + S_n \tag{9}$$

where S_n is the sample variance-covariance matrix.

The previous result can be used to combine market data with tweet data, assuming that the former represent "data" and the latter "prior information" in a Bayesian prior to posterior analysis.

To achieve this task we recall that, under a complete, fully connected graph, the expected value of the previous inverse Wishart is:

$$E(\Sigma|X) = T_n = (T_0 + S_n)/(\alpha + n) \tag{10}$$

and, therefore, the Bayesian estimator of the unknown variance covariance matrix, the a posteriori mean, is a linear combination between the prior (tweet) mean and the observed (market) mean.

When the graph G is not complete, a similar result holds locally, at the level of each clique and separator.

The previous results suggest to use the above posterior mean as the variance-covariance matrix of a complete graph on which to base model selection, thereby leading to a new selected graphical model, based on a "mixed" data source, that contains both financial and tweet data, in proportions determined by the quantities α and n. Model selection can be performed by maximising, rather than the likelihood, the Bayesian a posteriori probability. To achieve this task in an efficient way we will implement a Markov Chain Monte Carlo algorithm, following Giudici and Green (1999).

We now consider the issue of quantitative learning. In the context of systemic risk, a relevant quantity to be estimated is the partial correlation coefficient which, interpretationally, corresponds to the geometric mean between two regression coefficients in two differently directed multiple regression model. More formally:

$$\rho_{ijV} = \rho_{jiV} = \sqrt{a_{ijV} \cdot a_{jiV}}. \tag{11}$$

where a_{ijV} and a_{jiV} are, respectively, the regression coefficient of the multiple regression of X_i on all other V variables (including X_j) and the regression coefficient of the multiple regression of X_j on all other V variables (including X_i).

This interpretation of the partial correlation coefficient helps the construction of a novel contagion effect model, that can "modify" the probability of default of a financial institution with the effect of contagion from the institutions to which it is connected, in a level specified by the partial correlation coefficient.

For each node we assume to know the "idiosyncratic" probability of default of an institution, π_i, for example on the basis of the rating assigned by a credit rating agency, or from a credit scoring calculation, based on balance sheet data. From the probability of default we can derive, through the inverse Gaussian cumulative distribution function, the (idiosyncratic) credit score of the

corresponding institution, as follows:

$$Z_i^0 = \phi^{-1}(1 - \pi_i) \quad (12)$$

where π_i is the default probability of institution i and $1 - \pi_i$ is the corresponding survival probability.

We then assume that the idiosyncratic score of an institution can be modified through contagion, in a manner that depends on the credit scores of the neighbours, and on their partial correlations with i, as follows:

$$Z_i^{'} = \phi^{-1}(1 - \pi_0) - \sum_{j \in neigh(i)} a_{ij|rest} \phi^{-1}(1 - \pi_i) \quad (13)$$

where $a_{ij|rest}$ is the partial correlation coefficient between variables X_i and X_j given all the others (rest).

To interpret the previous assumption, consider the frequent case of positive partial correlations (which occur when banks are highly interrelated, as it occurs within the same country) and negative scores (which occur when default probabilities are less than 50%). In this case the idiosyncratic score increases through contagion and, therefore, the default probability increases too. This situation is illustrated in Figure 1 below.

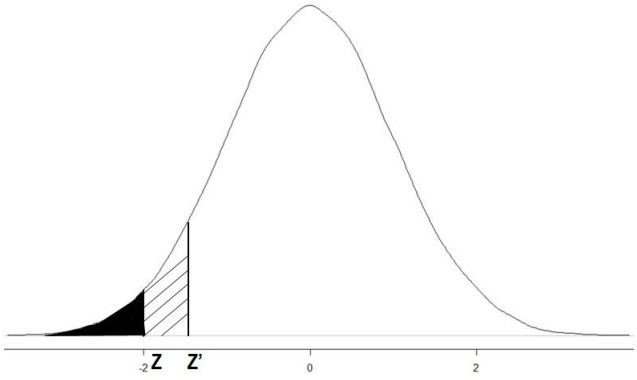


Figure 1: The impact of contagion on the probability of default: z is the credit score before contagion ans z' after contagion.

3. Application

In this section we consider the application of our proposed methodology. For reasons of information homogeneity we concentrate on a single market: the Italian banking system, characterised by a large number of banks, dominating the economy of the country, in a rapidly changing environment. We focus on large listed banks, for which there exists daily financial market data, that we would like to compare and integrate with tweet data.

Table 1 contains the list of banks that we consider, along with their total assets at the end of the last quarter of 2013 (in Euro), a measure of bank size. Banks are described by their stock market code (ticker).

Table 1 about here.

Bank Name	Ticker	Total Assets
UniCredit	UCG	926827
Intesa Sanpaolo	ISP	673472
Banca Monte dei Paschi di Siena	BMPS	218882
Unione di Banche Italiane	UBI	132433
Banco Popolare	BP	131921
Mediobanca	MB	72841
Banca popolare Emilia Romagna	BPE	61637
Banca Popolare di Milano	PMI	52475
Banca Carige	CRG	49325
Banca Popolare di Sondrio	BPSO	32349
Credito Emiliano	CE	30748
Credito Valtellinese	CVAL	29896

Table 1: List of considered listed Italian Banks

For each bank we consider the daily return, obtained from the closing price of financial markets, for a period of 148 consecutive days, from July 2013 to February 2014, as follows:

$$R_t = log(P_t / P_{t-1}) \quad (14)$$

where t is a day, $t - 1$ the day that preceeds it and P_t (P_{t-1}) is the corresponding closing price of that bank in that day.

For the same period, we have crawled Twitter, using the package TwitteR, available open source within the R project environment, and chosen all tweets that contain, besides one of the banks in Table 1, a keyword belonging to a financial taxonomy based on our knowledge of which balance sheet information may affect financial risk, as described in Table 2.

Table 2 about here

Assets	Liabilities	P& L
Liquidity	Deposits	Commissions
Corporate bonds	Customer deposits	Interest Margin
Government bonds	Allsale funding	Labour Costs
Loans	Interbank funding	Loans
Consumer loans	Capital	Loans losses
Derivatives	Equity	
	Shares	

Table 2: Initial proposed taxonomy analysis

Keywords in table 2 have been tested preliminarly to check which ones are the most effective in obtaining informative tweets. In table 3 we report only the relevant keywords, along with the relative frequencies.

Table 3 about here

Before extracting tweets, we have preliminarly filtered the most relevant financial twitterers, using the T-index methodology proposed in Cerchiello and Giudici (2015). Such methodology relies on an index that ranks sources according to the number of produced tweets,

Item	Frequency*100
Commissions	0.03
Labour costs	1.49
Deposits	0.08
Interbank	0.14
Management	28.58
Interest margin	4.91
Subsidiaries	0.99
Capital	35.67
Loan losses	0.73
Loans	10.11

Table 3: Final taxonomy

Item	Frequency*100	Average Sentiment
Commissions	0.03	2.67
Labour costs	1.49	3.21
Deposits	0.08	2.83
Interbank	0.14	2.19
Management	28.58	3.01
Interest margin	4.91	2.79
Subsidiaries	0.99	3.02
Capital	35.67	3.07
Loan losses	0.73	2.90
Loans	10.11	2.93

Table 4: Taxonomy proposed and descriptive sentiment analysis

and the corresponding re-tweets. The higher is the $T - index$, the stronger is the informative impact of a twitterer, because not only she/he produces many posts but also because they are highly shared among the community.

For a formal definition, given a set of n tweets of a tweeterer to which a count vector of the retweets of each tweet is associated, we consider the ordered sample of retweets $\{X_{(i)}\}$, that is $X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)}$, from which obviously $X_{(1)}$ ($X_{(n)}$) denotes the most (the least) cited tweet. Consequently the T index can be defined as follows:

$$T = max\{t : X_{(t)} \geq t\} \tag{15}$$

Once completed the preliminary phase as described above, each obtained tweet has been classified into a sentiment class, with categories ranging from 1 to 5. The higher the category, the more positive the sentiment (or value) that the tweet assigns to the bank under analysis, that is: 1=very bad, 2=bad, 3=neutral, 4=good and 5=very good. The sentiment classification has been carried out according to an appropriate classifier, trained on the data and employing a vocabulary of positive and negative Italian words adapted to the specific financial application under analysis. Such vocabulary is inspired by the famous Hu and Lu's opinion lexicon (first version described in Hu and Liu, 2004) that comprises around 6400 terms. In addition, several experiments and manual cross check have been carried out to improve the reliability and stability of the results. Moreover, since the total number of analyzed tweets is around 1000, thus easily manageable, the quality of the sentiment classification has been tested accurately comparing methods based on different versions of the vocabulary.

Table 4 describes the final employed taxonomy, along with the average sentiment associated to each keyword in our considered database. Here the sentiment scores are grouped by keywords, so that the average sentiment takes into account all the sentiment scores obtained for that specific word, regardless of the analysed bank.

Table 4 about here

For each bank we have then calculated a sentiment daily variation, that mimics market returns, as follows:

$$S_t = log(T_t/T_{t-1}) \tag{16}$$

where t is a day, $t-1$ the day that preceeds it, and T_t is the corresponding average daily sentiment on that bank for that day.

We now consider the application of our Bayesian model. In terms of prior parameters, we assume that $\alpha = n + 2$ and that T is a diagonal matrix, which implies zero a priori partial correlations.

In terms of structural learning, the selected model is the fully connected model: this is quite reasonable, in a national market that is fully integrated, with a strong country effect on bank risk.

Concerning quantitative learning, we report in Table 5, below the estimated partial correlations, obtained by model averaging them overthe most likely models (including, of course, the fully connected model). In Table 5 we also report, as a systemic risk measure for each bank, their weighted degree, calculated as the sum of all partial correlations, that expresses the intensity of the contagion.

Table 5 and, in particular, the weighted degree in the last row indicate, in a clear way, which are the most systemic banks: BPE, BP, followed by PMI and UBI: these are the four largest cooperative banks that are indeed linked to each other. The three largest (public) banks, UCG, ISP and MB, follow. Other smaller banks as well as the troubled MPS have a lesser degree. Note that Table 5 is also very useful to draw "stress test" analysis, such as: if UCG returns drop by 100 basis points, each of the other connected banks drop, on average, by 7 basis points, with a total impact on the system of 81 basis point. A similar drop in a smaller and relatively isolated bank, such as CVAL, causes an average drop of the other banks of only 3 basis points.

The above conclusions do not take bank size into account. However, it is very likely that the contagion effect of a bank on another also depends on the relationship between their sizes: the impact of a large bank, such as UCG, on a small bank, such as CE, is likely to

Bank	UCG	UBI	MB	ISP	CVAL	CE	BP	BPSO	PMI	BPE	BMPS	CRG
UCG	1.00	0.01	0.16	0.41	-0.11	-0.04	0.19	0.05	0.09	0.11	0.01	-0.01
UBI	0.01	1.00	0.20	0.03	-0.11	-0.04	0.26	-0.07	0.08	0.26	0.08	-0.03
MB	0.16	0.20	1.00	0.18	0.11	0.10	-0.05	0.10	-0.06	0.05	0.08	0.03
ISP	0.41	0.03	0.18	1.00	-0.09	0.02	-0.00	-0.01	0.13	0.01	0.01	0.00
CVAL	-0.11	-0.11	0.11	-0.09	1.00	-0.01	0.25	0.00	0.07	0.06	0.12	0.06
CE	-0.04	-0.04	0.10	0.02	-0.01	1.00	-0.02	0.09	0.08	0.14	-0.05	-0.04
BP	0.19	0.26	-0.05	0.00	0.25	-0.02	1.00	0.03	0.16	0.23	-0.01	-0.01
BPSO	0.05	-0.07	0.10	-0.01	0.00	0.09	0.03	1.00	0.15	0.04	0.00	0.05
PMI	0.09	0.08	-0.06	0.13	0.07	0.08	0.16	0.015	1.00	0.10	0.14	0.04
BPE	0.11	0.26	0.05	0.01	0.006	0.14	0.23	0.04	0.10	1.00	0.02	0.06
BMPS	0.01	0.08	0.08	0.01	0.12	-0.05	-0.01	0.00	0.14	0.02	1.00	0.11
CRG	-0.06	-0.03	0.03	0.00	0.06	-0.04	-0.01	0.05	0.04	0.06	0.11	1.00
Num. Links	11	11	11	11	11	11	11	11	11	11	11	11
Sum Par. Corr.	0.81	0.90	0.67	0.70	0.35	0.24	1.02	0.43	0.97	1.09	0.51	0.20

Table 5: Partial correlations and systemic risk measures based on the selected mixed graphical Gaussian model

be greater than what expressed by the weighted degree in Table 5.

To take size into account, we have inserted in the calculation of the contagion effect on the probability of default, in equation (1), a weight that is equal to the ratio of the total assets of the considered bank over the total market assets. Accordingly, Table 6 indicates the effect of contagion on the idiosyncratic PDs of the considered banks. The second and the third column of the table indicate the probability of default before and after contagion, and the corresponding percentage variation (Delta). For robustness purposes, we have also reported the same percentage variation assuming different values for the parameters of T: a common partial correlation of 0.8, rather than 0, which correspond to a more connected graph in the a priori twitter structure, and different values of *alpha*, which correspond to a higher weight for the twitter prior.

Bank	Contagioned PD	Delta	Delta (0.8%)	Delta ($3*(n+2)$)	Delta ($30*(n+2)$)
UCG	0.0064	+220	+220	+220	+220
UBI	0.0059	+195	+175	+200	+200
MB	0.0030	+50	95	+50	+50
ISP	0.0086	+330	365	+330	+335
CVAL	0.0055	-28	-1	-28	-28
CE	0.0073	-4	+20	-4	-4
BP	0.1450	+91	+66	+91	+91
BPSO	0.0025	+25	+75	+25	+25
PMI	0.1450	+108	+70	+109	+109
BPE	0.1340	+76	+80	+76	+78
BMPS	0.0024	+20	+5	+20	+20
CRG	0.0070	-8	+12	-8	-8

Table 6: Partial correlations and systemic risk measures based on the selected mixed graphical Gaussian model

From Table 6 note that the banks which are most vulnerable (most impacted by contagion) are the largest banks ISP, UCG as well as UBI, which is the most connected of the cooperative banks. In terms of robustness analysis, note that changing the a priori parameters do not change sensibly the results; this is especially true in terms of the correlation parameter. This indicates stability of the proposed model.

We remark that, in our opinion, the main aim of systemic risk analysis on banks is the understanding of the contagion effects, which is a quantitative learning problem. Selecting the best graphical model (the structural learning problem) is somewhat secondary, besides being more challenging, from a computational viewpoint. However, if there is a strong interest on searching the best structure, a computational approach, that relies on a penalised likelihood strategy, could be taken, for instance through a glasso approach (see e.g. Witten et al. 2011). While computationally appealing, the glasso has the disadvantage of arbitrariness in the choice of the penalty parameter λ. Such penalty deals with the level of complexity of the model, it is a regularization parameter used in the estimation of a sparse inverse covariance

matrix with a lasso (L1) penalty. Small λ corresponds to low levels of regularization, indeed if λ is set equal to 0, the graph results to be completely connected. On the opposite large values of λ corresponds to high penalization resulting in sparser, i.e. a less connected graph. Figure 2 below shows how the best chosen model changes considering different values of λ.

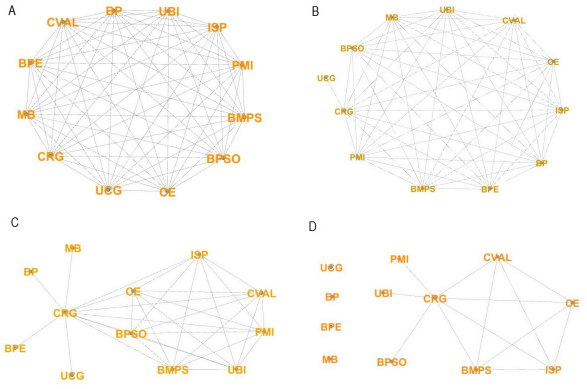


Figure 2: Selected Gaussian graphical models, with different Glasso parameters: A) $\lambda = 0$, B) $\lambda = 50$,C) $\lambda = 150$, D) $\lambda = 250$

From Figure 2 note that the selected graphical model sensibly depends on the choice of λ. While A and B correspond to a realistic highly interconnected situation, C and more so D correspond to less realistic sparse situations. It is worth mentioning the position of UCG according to λ values: in B, UCG starts isolating and such position is definitely confirmed in D. This happens coherently with the nature of UCG, that is the largest public Italian bank, the more international one and, therefore, the least exposed to national sources of stress.

4. Conclusions

In this paper we have shown how big data and, specifically, tweet data, can be usefully employed in the field of systemic risk modelling and, specifically, by means of graphical Gaussian models.

The paper shows how to combine tweet based systemic risk networks with those obtained from financial market data, using the a Bayesian model of data fusion and, correspondingly, a Bayesian model selection procedure.

We believe that our proposal can be very useful to estimate systemic risks and, therefore, to individuate the most vulnerable financial institutions. This because it integrates two different, albeit complementary, sources of information: market prices and twitter textual data.

Another important value of the model is its capability of including in systemic risk models institutions that are not publicly listed, using the tweet data component alone: a relevant advantage for banking systems where many banks are not listed, as it occurs in many European countries, for instance.

The model can be extended in several directions. A promising one could be to replace the inverse cumulative Gaussian link with an extreme value one, as in Calabrese and Giudici (2015) so to focus more the analysis on tail events.

References

[1] Acharya VV, Pedersen LH, Philippon T, Richardson M (2010) Measuring systemic risk. Working paper, Federal Reserve of Cleveland.

[2] Adrian T, Brunnermeier MK (2009). CoVaR. Technical report, Princeton University.

[3] Benoit S, Colliard JE, Hurlin C,Perignon C (2015). Where the Risks Lie: A Survey on Systemic Risk. HEC Paris Research Paper No. FIN–2015–1088.

[4] Billio M, Getmansky M, Lo A, Pelizzon L (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sector. *Journal of Financial Economics*, **104(3)**, 535–559.

[5] Bollen J, Mao H, Zeng X (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, **2(1)**, 1-8.

[6] Bordino I, Battiston S, Caldarelli G, Cristelli M, Ukkonen A, Weber I (2012). Web search queries can predict stock market volumes. *PloS one*, **7(7)**, e40014.

[7] Boss M, Elsinger H, Summer M, Thurner S (2004). Network topology of the interbank market. *Quantitative finance*, **4(6)**, 677-684.

[8] Brownlees CT, Engle RF (2011). Volatility, correlation and tails for systemic risk measurement. Technical report, New York University.

[9] Brunnermeier M, Oehmke M (2012). Bubbles, Financial Crises, and Systemic Risk. NBER Working Papers 18398, National Bureau of Economic Research.

[10] Calabrese R., Giudici P. (2015). Estimating bank default with generalised extreme value regression models. *Journal of the Operational Research society*, 1–10, doi:10.1057/jors.2014.106.

[11] Carlin B.P., Louis T.A. (2000). Bayes and Empirical Bayes Methods for Data Analysis (2nd ed.). Chapman & Hall/CRC.

[12] Casella G. (1985). An Introduction to Empirical Bayes Data Analysis. *American Statistician (American Statistical Association)* **39(2)**, 83-87.

[13] Cerchiello P, Giudici P (2016). Conditional graphical models for systemic risk estimation (2016) To appear in Expert systems with applications. 10.1016/j.eswa.2015.08.047 Vol. 43, pp. 165-174.

[14] Cerchiello P, Giudici P (2015). How to measure the quality of financial tweets. Quality and Quantity. DOI 10.1007/s11135-015-0229-6.

[15] Choi H, Varian H (2012). Predicting the present with google trends. *Economic Record*, **88(s1)**, 2–9.

[16] Dawid AP, Lauritzen SL (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, **21**, 1272–317.

[17] Diamond DW, Dybvig PH (1983). Bank runs, deposit insurance, and liquidity. *Journal of Political Economy* **91(3)**, 401-419, Reprinted (2000) *Fed Res Bank Mn Q Rev*, **24(1)**, 14-23.

[18] Eisenberg L, Noe T (2001). Systemic risk in financial systems. *Management Science* **47(2)**, 236–249.

[19] Feldman, R.(2013). Techniques and applications for sentiment analysis. *Commun. ACM*, **56(4)**, 82-89.

[20] Giudici, P. (2001). Bayesian data mining, with application to financial benchmarking and credit scoring. *Applied stochastic models in business and industry*, **17**, 69–81.

[21] Giudici P, Green PJ (1999). Decomposable graphical Gaussian model determination. *Biometrika*, **86**, 785–801.

[22] Giudici P, Spelta A (2015). Graphical network models for international financial flows (2015). Journal of Business and Economic Statistics. DOI 10.1080/07350015.2015.1017643.

[23] Huang X, Zhou H, Zhu H (2011). Systemic risk contribution. Technical report, Board of Governors of the Federal reserve System.

[24] Hu M., Liu B. (2004). Mining and summarizing customer reviews. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery Data Mining (KDD-2004, full paper), Seattle, Washington, USA.

[25] Idier J, Lame' G, Mesonnier JS (2013). How useful is the marginal expected shortfall for the measurement of systemic exposure? a practical assessment. Working paper series, 1546, European Central Bank.

[26] Lauritzen SL (1996). Graphical models. *Oxford University Press*.

[27] Mantegna RN(1999). Hierarchical structure in financial markets. *The European Physical Journal B*, **11**, 193–197.

[28] Merton RC (1974). On the pricing of corporate debt: the risk sructure of interest rates. *Journal of Finance*, **2**, 449–471.

[29] Nymand-Andersen (2016), Big data: The hunt for timely insights and decision certainty Per. *IFC Working Papers*, **14**.

[30] Onnela JP, Kaski K, Kertesz J (2004). Clustering and information in correlation based financial networks. *European Physical Journal B*, **38**, 353–362.

[31] Sheldon G, Maurer M (1998). Interbank Lending and Systemic Risk: An Empirical Analysis for Switzerland. *Swiss Journal of Economics and Statistics (SJES)*, **134(4)**, 685–704.

[32] Upper C, Worms A (2004). Estimating bilateral exposures in the German market: is there a danger of contagion? *European economic review*, **48(4)**, 827–849.

[33] Tumminello M, Aste T, Di Matteo T, Mantegna RN (2007). Correlation based networks of equity returns sampled at different time horizons. *European Physical Journal B*, **55(2)**, 209–217.

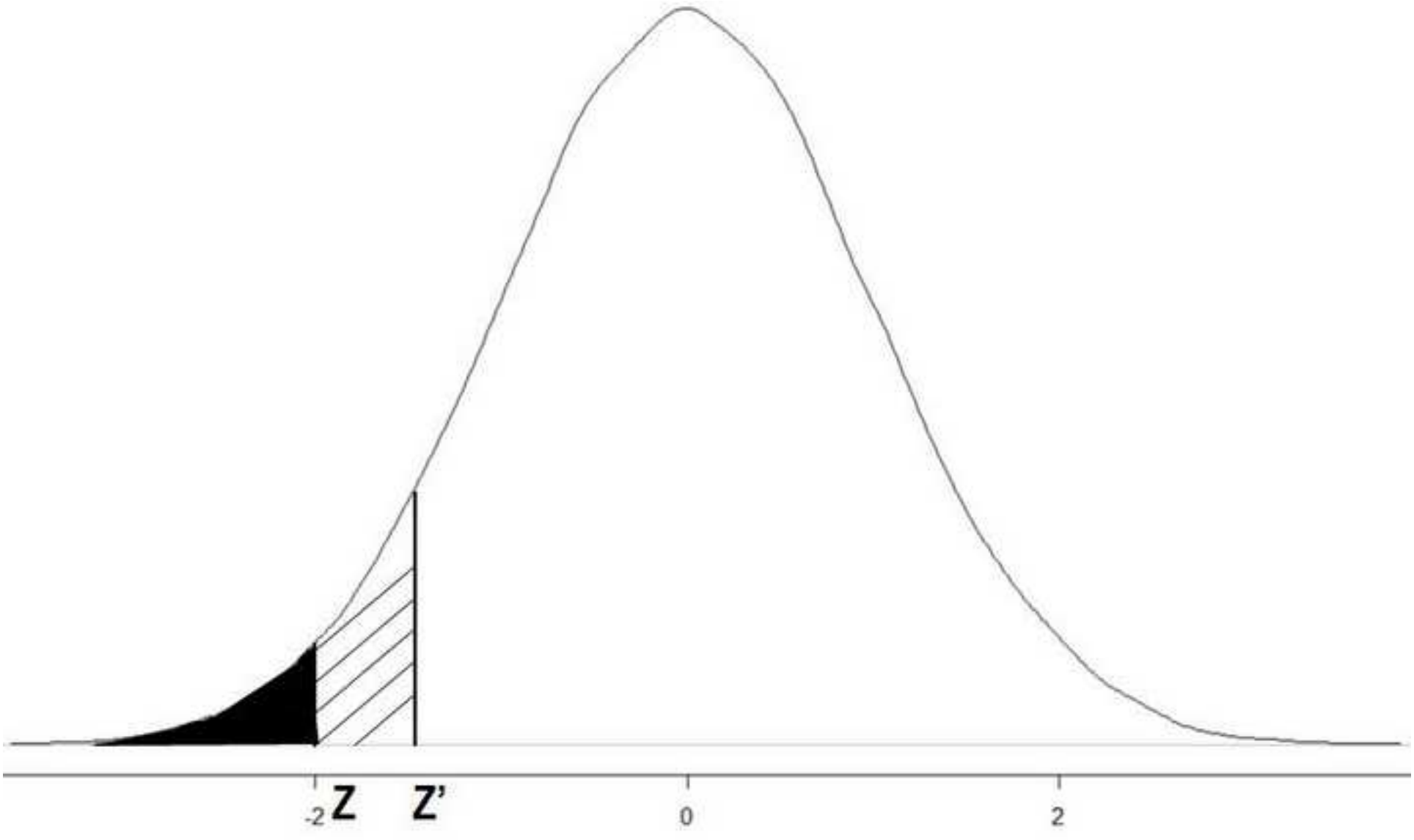
[34] Vasicek O. A. (1984). Credit valuation. KMV corporation, March.

[35] Whittaker J (1990). Graphical models in applied multivariate statistics. *Wiley Publishing*.

[36] Witten W, Friedman J and Simon N (2011). New insights and faster computations for the graphical lasso. Journal of Computational and Graphical Statistics, Volume 20, Number 4, pp. 892–900.

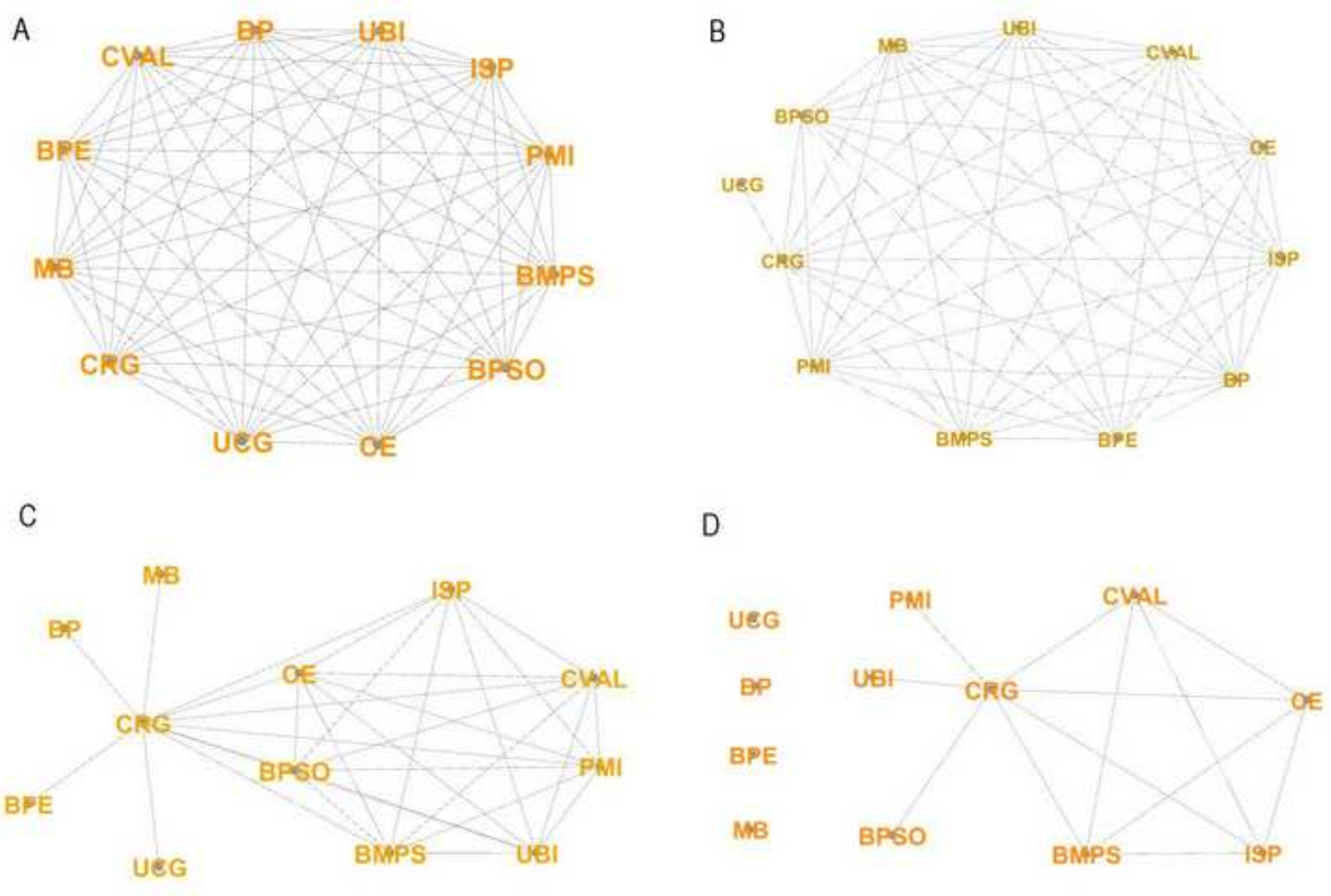
Figure

Click here to download high resolution image



Figure

Click here to download high resolution image



Paola Cerchiello is researcher of Statistics at the Department of Economics and Management of the University of Pavia.

Her research activity is mainly devoted to the study of statistical for ordinal and qualitative models in economics and finance. From an applied viewpoint, she focuses on text data analysis, operational and reputational risk, sentiment analysis and teaching quality evaluation. She has published around 30 papers in scientific international journals and has an H-index of 6 (calculated by Google scholar).

Paolo Giudici is Full Professor of Statistics at the Department of Economics and Management of the University of Pavia.

His research activity concerns statistical models for economics and finance and, in particular, data mining models, applied bayesian analysis, graphical association models and models for credit and operational risk. He has published 57 papers in scientific international journals, one research book on Applied data mining, and has an H-index of 19 (calculated by Google scholar).

***Biography of the author(s)**

Click here to download Biography of the author(s): Bio_Nicola.docx

Ph.D student in Data Science,
research interests in Text analysis, Network analysis and Machine Learning
Work experience in Management Consulting at Roland Berger Strategy Consultants
M.Sc. in Nuclear Engineering at Polytechnic of Milan

***Photo of the author(s)**

Click here to download high resolution image

*Photo of the author(s)

Click here to download high resolution image

***Photo of the author(s)**

Click here to download high resolution image

Source

Click here to download Source Files - Latex or Word: neurocomputing rev3.tex