Arcada Working Papers 2/2015
ISSN 2342-3064
ISBN 978-952-5260-61-8

Bank Networks from Text:

Interrelations, Centrality and Determinants

Samuel Rönnqvist[i], Peter Sarlin[ii]

Abstract

In the wake of the still ongoing global financial crisis, interdependencies among banks have come into focus in trying to assess systemic risk. To date, such analysis has largely been based on numerical data. By contrast, this study attempts to gain further insight into bank interconnections by tapping into financial discourse. We present a text-to-network process, which has its basis in co-occurrences of bank names and can be analyzed quantitatively and visualized. To quantify bank importance, we propose an information centrality measure to rank and assess trends of bank centrality in discussion. We also analyze determinants of information centrality to better understand driving factors behind the importance of banks in the network. For qualitative assessment of bank networks, we put forward a visual, interactive interface for better illustrating network structures. We illustrate the text-based approach on European Large and Complex Banking Groups (LCBGs) during the ongoing financial crisis by quantifying bank interrelations from discussion in 1.3M news articles, spanning the years 2007 to 2013.

Keywords: text analytics, network analysis, systemic risk

i Department of Information Technologies / TUCS – Turku Centre for Computer Science, Åbo Akademi University, Finland, sronnqvi@abo.fi

ii RiskLab Finland, Arcada University of Applied Sciences, Finland; Center of Excellence SAFE, Goethe University Frankfurt, Germany; Department of Economics, Hanken School of Economics, Finland; peter@risklab.fi

1 INTRODUCTION

The global financial crisis has brought several banks, not to say entire banking sectors, to the verge of collapse. This has not only resulted in losses for investors, but also costs for the real economy and welfare at large. Considering the costs of banking crises, the recent focus of research on financial instabilities is well-motivated. First, real costs of systemic banking crises have been estimated to average at around 20-25% of GDP (e.g., [16, 21]). Second, data from the European Commission illustrate that government support for stabilizing banks in the European Union (EU) peaked at the end of 2009. The support amounted to €1.5 trl, which is more than 13% of EU GDP. The still ongoing financial crisis has stimulated a particular interest in systemic risk measurement through linkages, interrelations, and interdependencies among banks. This paper advances the literature by providing a novel measure of bank linkages from text and bank importance through information centrality.

Most common sources for describing bank interdependencies and networks are based upon numerical data like interbank asset and liability exposures, and co-movements in market data (e.g., equity prices, CDS spreads, and bond spreads) (see [14]). While these direct and indirect linkages complement each other, they exhibit a range of limitations. Even though in an ideal world bank networks ought to be assessed through direct, real linkages, interbank data between banks' balance sheets are mostly not publicly disclosed. In many cases, even regulators have access to only partial information, such as lack of data on pan-European bank linkages despite high financial integration. Market price data, while being widely available and capturing other contagion channels than those in direct linkages between banks [1], assume that asset prices correctly reflect all publicly available information on bank risks, exposures and interconnections. Yet, it has repeatedly been shown that securities markets are not always efficient in reflecting information about stocks (e.g. [22]). Further, co-movement-based approaches, such as that by Hautsch et al [19], require large amounts of data, often invoking reliance on historical experience, which may not represent the interrelations of today. Also, market prices are most often contemporaneous, rather than leading indicators, particularly when assessing tail risk. It is neither an entirely straightforward task to separate the factors driving market prices in order to observe bilateral interdependence [12].

Big data has emerged as a central theme in analytics during the past years. Research questions of big data analytics arise not only from massive volumes of data, or speeds at which data are constantly generated, but also from the widely varying forms, particularly unstructured textual data, that in themselves pose challenges in how to effectively and efficiently extract meaningful information [17]. This paper treats the text mining aspect, as it proposes an approach to relationship assessment among banks by analyzing how they are mentioned together in financial discourse, such as news, official reports, discussion forums, etc. The idea of analyzing relations in text is in itself simple, but widely applicable. It has been explored in various areas; for instance, Özgür et al. [31] study co-occurrences of person names in news, and Wren et al. [30] extract biologically relevant relations from research articles. These approaches can be used to construct social or biological networks, using text as the intermediate medium of information. Our contribution lies in proposing this text-based approach to the study of

bank interrelations, with emphasis on analysis of the resulting bank network models and ultimately quantifying a bank's importance or centrality.

Our approach may be compared to the above discussed, more established ways of quantifying bank interdependence, such as interbank lending and co-movement in market data. While not measuring direct interdependence, it has the advantage over interbank exposures by relying upon widely available data, and over co-movements in market data by being a more direct measure of an interrelation. On the other hand, our approach serves to shed light on banks' relationships in the view of public discussion, or of information overall, depending on the scope of textual data. It may serve as a way of tapping into the wisdom of the crowd, while offering a perspective different from previous methods, especially considering the presence of rich, embedded contextual detail. Rather than an ending point, this sets a starting point from which further study may focus more extensively on the context of occurrences and more sophisticated semantic analysis. This allows to better understand factors driving interrelations, and overall centrality.

We assess European Large and Complex Banking Groups (LCBGs) using the text-based approach for quantifying bank interrelations from discussion in the news. A co-occurrence network is derived from 1.3M articles, spanning the years 2007 to 2013 in the Reuters online news archive. Beyond only quantifying bank interrelations, we also provide means for quantitative and qualitative assessment of networks. To support quantification of bank importance, we propose an information centrality measure to rank and assess trends of bank centrality in discussion. Rather than a common shortest-path based centrality measure, information centrality captures effects that might propagate aimlessly by accounting for parallel paths. To support a qualitative assessment of the bank networks, we put forward a visual, interactive interface for better illustrating network structures. This concerns not only an interface to network models, but also an interactive dashboard to better communicate quantitative network measures.[1]

The co-occurrence network illustrates relative prominence of individual banks, and segments of more closely related banks. The systemic view acknowledges that the centrality of a bank in the network is a sign of importance, and not necessarily its size (cf. too central to fail by Battistone et al. [7]). The dynamics of the network, both local and global, reflect real-world events over time. The network can also be utilized as an exploratory tool that provides an overview of a large set of data, while the underlying text can be retrieved for more qualitative analysis on relations.

To better understand what drives information centrality, and how it ought to be interpreted, we explore determinants of the centrality measure. We investigate a large number of bank-specific risk drivers, as well as country-specific macro-financial and banking sector variables, as well as control for variables measuring bank size. Further, we also assess the extent to which information centrality explains banks' risk to go bad, and compare it to more standard measures of size. Even though bank size is a key factor explaining information centrality, we show that centrality is not a direct measure of

1 The interactive implementations are available online at: http://risklab.fi/demo/textnet/

vulnerability. This implies that the centrality measure is not biased by the nature of business activities or models, which potentially impacts bank vulnerability (e.g., asset size or interbank-lending centrality). Rather, while not being a narrow, direct measure of interconnectedness, we are capturing broad importance of a bank in terms of information connectivity in financial discourse from a wider perspective. Yet, while the rich nature of textual data provides means for more specific queries in defining interrelationships and other potentially interesting details on banks, its interpretation by computational methods is often challenging. To this end, we also discuss different ways of analyzing text-based networks, laying forward some ideas on future directions in the study of them.

The following section explains the data and methods we use to construct and analyze bank networks from text, whereas Section 3 discusses the results of the experiments on textual data, including both qualitative and quantitative analysis. Before a concluding discussion on text-based networks, Section 4 assesses determinants of information centrality.

2 BANK NETWORKS FROM TEXT: DATA AND ANALYSIS

This section provides a discussion of the text-to-network process, both generally and from the viewpoint of the study in this paper. First, we detail the particular text data and choice of banks to be studied. Having established this, we turn to the process of text analysis and construction of bank co-occurrence networks. This is followed by discussion on the analysis of such networks, including both quantitative and qualitative analysis.

2.1 Data and target banks

Through digitized economic, social and academic activities, we are having access to ever increasing amounts of textual data. While vast amounts of textual data are readily available, there is nothing that assures increases in precision and quality of data. Analytics of big data is increasingly a search for needles in a haystack, where choices in data source, collection methods as well as pre-processing setups all need to be carefully directed in order to pick up desired signals. Likewise, when tapping into financial discourse, one needs to clearly narrow the context of collected data and targeted entities of interest, beyond the choice of data source.

The text data we use in this paper is newly collected from Reuters online news archive. News text presents a rather formal type of discourse, which eases interpretation of extracted relations, as opposed to more free-form, user-generated online discussion. We focus on major consumer banks within Europe, classified by the European Central Bank [2] as Large and Complex Banking Groups (LCBGs), of which 15 are also classified as Globally Systemically Important Banks (G-SIBs) by the Financial Stability Board [1]. See Appendix A for a list of LCBGs and G-SIBs and the naming convention used in this paper. The period of study is 2007-2013, for which the news archive contains 6.4M

articles. We base our analysis on a 20% random sample of articles comprising of 1.3M article.

The text analysis is based on detecting mentions of bank names in the articles. We look at a set of 27 banks: 5 British, 5 French, 4 German, 4 Spanish, 3 Dutch, 2 Italian, 2 Swiss, 1 Swedish and 1 Danish bank. In order to mitigate a geographical sampling bias, we use the U.S. edition of the Reuters news archive, as no single European edition is available, but rather national editions for only the largest countries.

The chart in Figure 1 provides an overview of the trends in total news article volume, as well as the volume of bank name occurrences. Out of all articles, 6% mention any of the targeted banks, on average. The volume is relatively low in the beginning of 2007, i.e., the start of the archive. Mentions of banks reach a peak in early 2008, but returns to a stable level all through 2013.

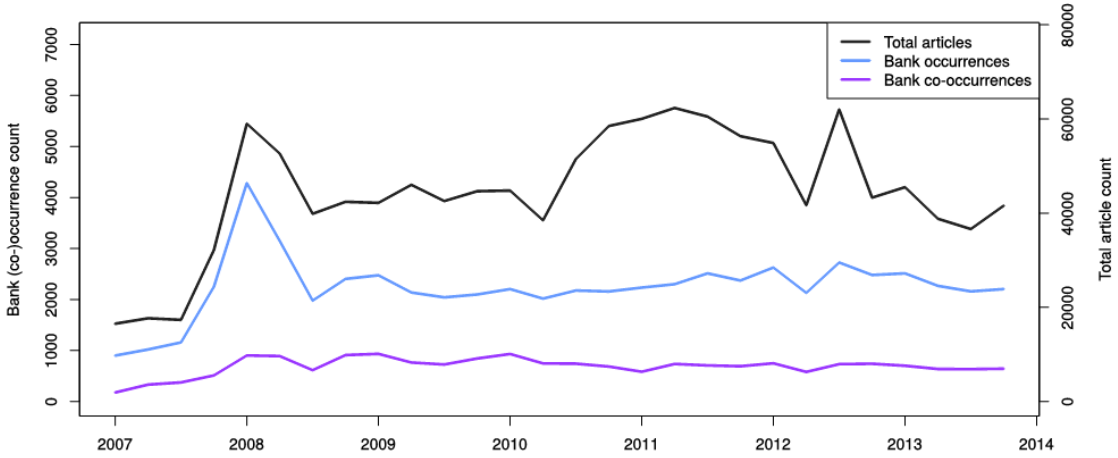


Figure 1. Volumes of all news articles and bank name occurrences over time.

2.2 From text to bank networks

With plain text as a starting point, and relationship assessment as an objective, we analyze co-mentions in financial discourse. Extracting occurrences and co-occurrences from text is the initial step. The relationships are constituents of co-occurrence networks, whose properties can be assessed through both quantitative and visual analysis. Figure 1 provides an overview of the process of transforming text into network models that lend themselves to analysis.

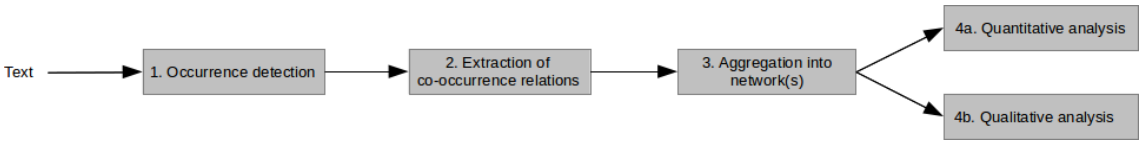


Figure 2. Text-to-network process: (1) Occurrences of bank names are detected in source text, (2) pair-wise co-occurrence relations are extracted between occurrences within a context, and (3) relations aggregated over a time interval form a co-occurrence network. A resulting network can be analyzed with (4a) quantitative measures capturing some interesting features, and (4b) qualitative analysis through visual exploration of the network, its neighborhoods, and connectivity of individual nodes.

To construct the network we scan the text for occurrences of bank names to detect and register mentions of those banks. Scanning is performed using patterns (regular expressions), manually designed and tested to match with as high accuracy as possible. Generally, the use of manually designed patterns for information extraction in text tend to have high accuracy but lower recall, but we expect that the reasonably standardized form of discourse we use should mitigate a loss in recall.

A co-occurrence relation is formed by two bank names occurring in the same context. Multiple occurrences of a single bank are counted only once per context, ignoring meaningless repetitions, but an occurrence may participate in multiple relations. In the present case, we define the scope of the context as a 400-character window, whereas a wider scope would require less data but increase noise. A context containing two or more banks yields one or more pair-wise co-occurrence relations. Contexts with more than 5 banks are disqualified, as they are likely to be listings that would result in marginally meaningful relations. These pre-processing design decisions should be adjusted and tested for each new data source, to obtain less noisy results.

Aggregated into a network, the extracted relations can be studied using methods for analysis of complex networks. In the network, banks form nodes (or vertices), and aggregated co-occurrence relations form links (or edges). Each link is weighted according to the aggregated count of co-occurrences, over a certain time interval. To extract meaningful quantitative measures of co-occurrence networks, measures designed for weighted networks need to be used. Nevertheless, most conventional network analysis methods are designed for binary (unweighted) networks only [25], which calls for some form of transformation of the network if these measures are to be used, such as by filtering out very weak connections. While unfiltered networks are more sensitive to noise when using binary measures, low-frequency co-occurrences may be of particular interest, as they are more likely to represent novel information. In order not to lose detail, it is highly motivated to use weighted networks and measures that account for link weights.

Although quantitative analysis of networks provides means to better understand overall properties of networks, they as any aggregate measure most often lack in detail. Hence, visual representations provide ample means for not only detailed analysis of the underlying constituents of the networks, but also further details as demanded. In the following subsection, we further discuss both quantitative measurement of network properties and visualization as a support in their analysis.

2.3 Network analysis

Network models are commonly rather complex and rich in information. They can be analyzed in many different ways to gain insight into the nature of the underlying phenomenon, the bank connectivity landscape in our case. We first discuss analysis of the networks at a global, descriptive level, to describe properties of the co-occurrence networks through common network measures. Later, we concentrate on the concept of centrality and a few ways of quantifying it in our type of network, with the study of systemic risk in mind. Finally, we discuss network visualization as a means for interactive exploration.

2.3.1 Global properties

A commonly cited property of real-world networks is that the average distance between nodes is very small relative to the size of the network, lending them the name ”small-world” networks [29]. Short distances have a functional justification in most types of network, as it increases efficiency of communication, while there also is a general tendency towards short average distances among non-regular networks. These networks have varying degree, i.e., number of links per node, the distribution of which is a typical way of profiling empirical networks. Networks that have evolved through natural, self-organizing processes, such as communications, social, biological and financial networks, tend to exhibit degree distributions that follow a power law. These so-called scale-free networks evolve through processes of preferential attachment, where the likelihood of a node receiving a new link is proportional to its current degree [4].

Jackson & Rogers [20] distinguish two archetypes of natural networks, described by power-law degree distributions and exponential degree distributions, respectively. They argue that, in fact, empirical networks generally exhibit hybrid distributions, between power-law and exponential, as they are formed through mixed processes of preferential attachment and attachment with uniform probability. The latter process still generates highly heterogeneous exponential distributions, as established nodes have greater chance over time at growing well embedded into the network. By either process, some nodes are bound to be more influential than others, and mapping the levels of influence in the system is our main interest. To profile the co-occurrence networks, the average shortest paths and degree distributions can indicate how small-world and scale-free they are. In the latter case, as we are interested in accounting for the link weighting, we study the distribution of strength, i.e., weighted degree calculated as the sum of weights per node (as [6] propose).

2.3.2 Centrality

Following the initial profiling of the whole network, we turn the focus towards the concept of node centrality. A central node holds a generally influential position in a network; a centrally located bank is likely to be systemically important, as it stands to affect a large part of the network directly or indirectly in case of a shock (negative or positive). There is, however, a range of ways to quantify centrality, the most common

measures being degree centrality (i.e., fraction of nodes directly linked) and the shortest-path-based closeness centrality and betweenness centrality. We adapt degree centrality to our weighted networks, by using strength as a direct measure of centrality. Closeness and betweenness centrality can also incorporate link weight into the calculation of shortest path, by means of Dijkstra's shortest-path algorithm [18] that interprets weights as distances between nodes. Since co-occurrence networks represent tighter connections (i.e., more co-occurrences) by higher weights, it is necessary to invert the weights before calculation, as proposed by [24].

Borgatti [11] points out that a common mistake in the study of network centrality is to neglect to consider how flow in the system is best modeled. The common shortest-path based centrality measures make implicit assumptions that whatever is passing from a node to the surrounding network does so along optimal paths, such as in routing networks of goods and targeted communication. Arguably, a more realistic intuition for influence of a node, in cases where effects might propagate aimlessly, such as any type of contagion, is one that accounts for parallel paths that may exist.

Along these lines, we study a closeness centrality measure that models the flow of information in such a manner, called information centrality [27] (also known as current flow closeness centrality [13]). Information centrality, which seeks to quantify the information that can pass from a node to the network over links whose strength determine level of loss in transmission, is defined as

$$I(i) = \frac{n}{nC_{ii} + \sum_{j=1}^{n} C_{jj} - 2\sum_{j=1}^{n} C_{ij}}$$

where n is the number of nodes and the weighted pseudo-adjacency matrix is defined as

$$C = B^{-1},\ B_{ij} = \begin{cases} 1 + S(i),\ if\ \ i = j \\ 1 - w_{ij},\ \ otherwise \end{cases}$$

where w is link weight (0 for unlinked nodes) and $S(i)$ is strength of node i.

Centrality as a measure of a node's relative importance is interesting, yet changes in centrality adds another dimension. We study networks of quarterly cross sections of the data, in order to calculate and compare centralities over time.

When the data is split into smaller parts less frequent parts will inevitably become disconnected from the main network component. Information centrality can be quite sensitive to the resulting fluctuations in component size, while the more central nodes start to correlate strongly. We propose a method to stabilize the centrality measurement by applying Laplace smoothing to the link weights before calculation of information centrality. The weight of each existing link is increased by a small constant (e.g. 1.0), while links are added between all other nodes and weighted by the same constant. Formally, $W'_{ij} = w_{ij} + \alpha$, where w_{ij}=0 if i and j are not connected. The reasoning is that operating on a limited sample of links, we want to discount some probability for unobserved links (between known nodes), to lessen the influence that the difference between non-occurring (unobserved) links and single-occurrence links has on centrality.

This type of additive smoothing has similarly been applied in language modeling [15], but is generally applicable to smoothing of categorical data.

2.3.3 Visual analysis

While quantitative network analysis plays a vital role in measuring specific aspects of interest in a precise and comparable fashion, network visualization can provide useful overview and exploratory capabilities, communicating general structure as well as local patterns of connectivity. The visual analytics paradigm aims at supporting analytical thinking through interactive visualization, where interaction is the operative term. Through a tight integration between the user and the data model, users are enabled to explore and reason about the data. In the case of our dynamic networks, interaction capabilities for navigating between cross sections and further exploring network structure provide a setting for qualitative analysis of the information-rich models.

Force-directed layouting is often used to apply spatialization of network nodes, that is, to place the nodes in a way that overall approximates node distances to their corresponding link strengths, thereby seeking to uncover the structure of the network in terms of more and less densely connected areas and their relation. Still, force-directed layouts quickly turn uninformative or ambiguous as the networks become too dense, including cases of weighted networks with few strong but many weak connections. Although network visualization with force-directed layouting often does not scale well to analysis of big networks, it still can be a useful tool when used properly. In the case of our bank co-occurrence network it produces decent visualizations for cross sections of the data set, while stricter filtering of co-occurrences will produce a sparser network that is less cluttered. We use the D3 force algorithm [3] for layouting.

The dynamics of the network can be studied by visualizing cross-sectional networks in a series, where the positioning is initialized by the previous step and optimized according to the current linkage, as to provide continuity that helps in the visual exploration of network evolution. User interaction plays a vital role not only by allowing to navigate across time, but also by allowing interaction with the positioning algorithm, letting the user acquire a more direct understanding of the structures and details in the data. Force-directed layouting on more densely linked networks generally finds a locally optimal positioning out of a large number of comparably good solutions. Interaction that lets the user drag nodes to reposition them and a force-directed algorithm that helps to counter-optimize the positioning immediately afterwards gives rise to a collaborative, exploratory way of working with and understanding the data.

Nevertheless, the best setting for visual analysis might be one that combines with quantitative analysis, encoding them visually. For instance, centrality measures can be encoded by node size to enhance the communication of structure provided by the network visualization, which can use force-directed layouting or other more regularly structured layouts. Hence, information centrality might be considered as a means to encode node size.

3 CENTRALITY: QUANTITATIVE AND VISUAL ANALYSIS

This section describes the co-occurrence networks from both a viewpoint of quantitative measures and exploratory visualization. With the assessment of network measures as a starting point, we describe network properties in general and information centrality in particular. Then, we turn to visual analysis of the networks and their constituents.

3.1 Quantitative analysis

The volume of bank occurrences is remarkably stable across time, apart from a peak centered around 2008Q1. At that time the peak in total article volume coincides with a peak in occurrence volume, unlike later during the studied time span when occurrence seems unaffected by fluctuating article volume. Interestingly, the 2008 surge in occurrences does not translate into a rise in co-occurrences (or strength), i.e., even though banks are more discussed at the time prior to the outbreak of the crisis, they are not discussed more in close connection to each other.

From these aggregated counts, we continue by studying the data as a network. As discussed in Section 2.3, empirical networks are typically profiled through measures describing certain global properties. The average distance, in terms of number of links, between nodes in the co-occurrence networks are certainly small, at 1.1-1.3, and would justify calling them 'small-world' networks. However, with weighted links, a measure of average distance becomes hardly interpretable. While it is clear that our networks are very tightly connected, the strength distribution depicts the relative differences in node connectivity. Many empirical networks exhibit power-law distributed degree or strength, as a sign of evolution through preferential attachment. Figure 3 shows the cumulative strength distribution of the aggregated network for the years 2007-2013, as well as a closely fitted exponential function that hints that our network is exceedingly a product of evolution through uniform attachment. Still, we are able to partially fit power-law functions to the distribution, as the figure highlights with straight lines, which could indicate a hybrid model with a weak preferential attachment component as well. The strength distribution illustrates the high heterogeneity of connections in the network, i.e., some banks are much more associated in discussion than others. However, in order to gain a deeper understanding of a bank's importance to the wider network, we need to look beyond immediate connections as measured by degree/strength distribution or degree/strength centrality (proportional to co-occurrence volume), namely we need to look at information centrality.

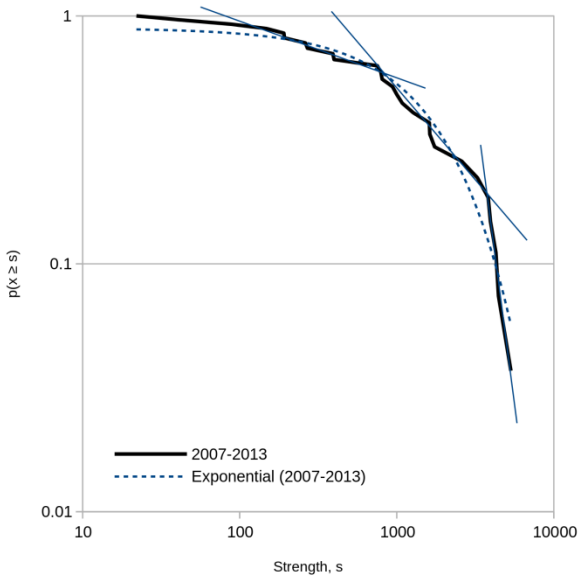


Figure 3. Cumulative strength distribution (weighted degree) of bank co-occurrence network during 2007--2013, showing probability p over node strengths. Dashed line is a fitted exponential function. Solid straight lines indicate locally fitting power-law functions.

We study information centrality for each node over time, using different levels of Laplace smoothing. Figure 4 plots the information centrality values, with a number of example banks highlighted in color. Information centrality without smoothing exhibits a number of peaks of comparable magnitude. Compared to the case of light smoothing (α=0.1), which levels all peaks except in 2008Q4 (crisis breakout), it appears that most peaks of unsmoothed information centrality are in fact meaningless artifacts of changing network size. Further, stronger smoothing (α=1.0) does not have as strong an effect on artifact peaks, but it does help to even the distribution of banks on the information centrality scale, so that fewer banks flock at the top.

The trends of individual banks generally follow the movements of the cross section closely, as increased connectivity in parts of the network strongly affects the rest, since the co-occurrence network is generally very tightly connected. Individual centrality relative to the cross section is generally quite stable. Nevertheless, some changes can be observed that might reflect real-world events. For instance, ABN AMRO has relatively high information centrality in 2007 that decreases afterwards. Royal Bank of Scotland is the most central bank in 2007-2008, whereas it later on is overtaken by Barclays and Deutsche Bank. To illustrate the information centrality ranking between banks in more detail, Figure 5 shows all values as of 2013Q4.

In the smoothed information centrality plots both 2008Q1 and 2008Q4 exhibit peaks. In the first quarter, the peak coincides with the peak in bank occurrence. The fact that co-occurrence stays flat during the same time indicates that the change in information centrality is not due to generally strengthened connections, but rather due to change in topology. The peak in the fourth quarter likewise hints at topological shifts following the crisis outbreak. Some slight upward movements can also be hinted around year 2012.

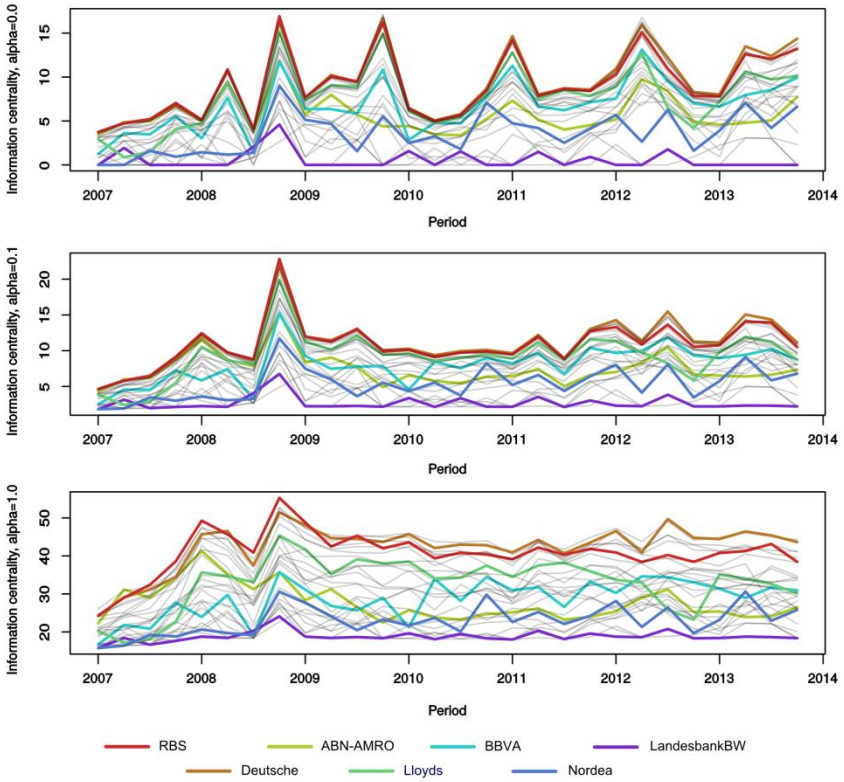


Figure 4. Information centrality for banks over time. The charts show different levels of smoothing: none (α=0.0), little (α=0.1) and moderate (α=1.0). A few example banks are highlighted (bank labels are described in Appendix A).

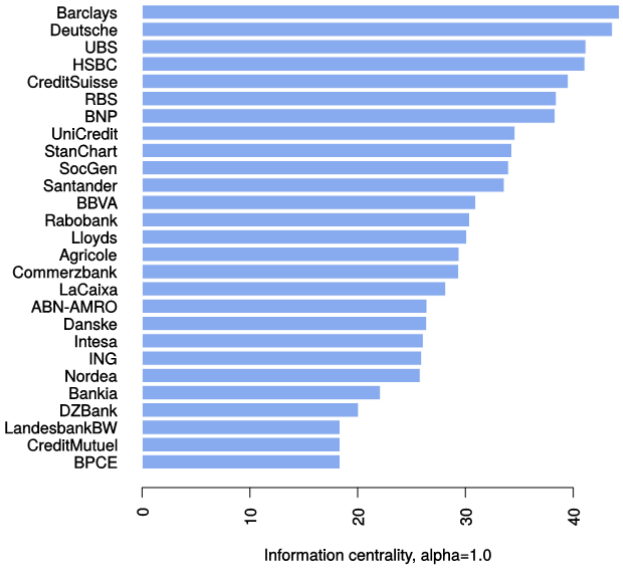


Figure 5. Information centrality ranking for all banks in 2013Q4 (bank labels are listed in Appendix A).

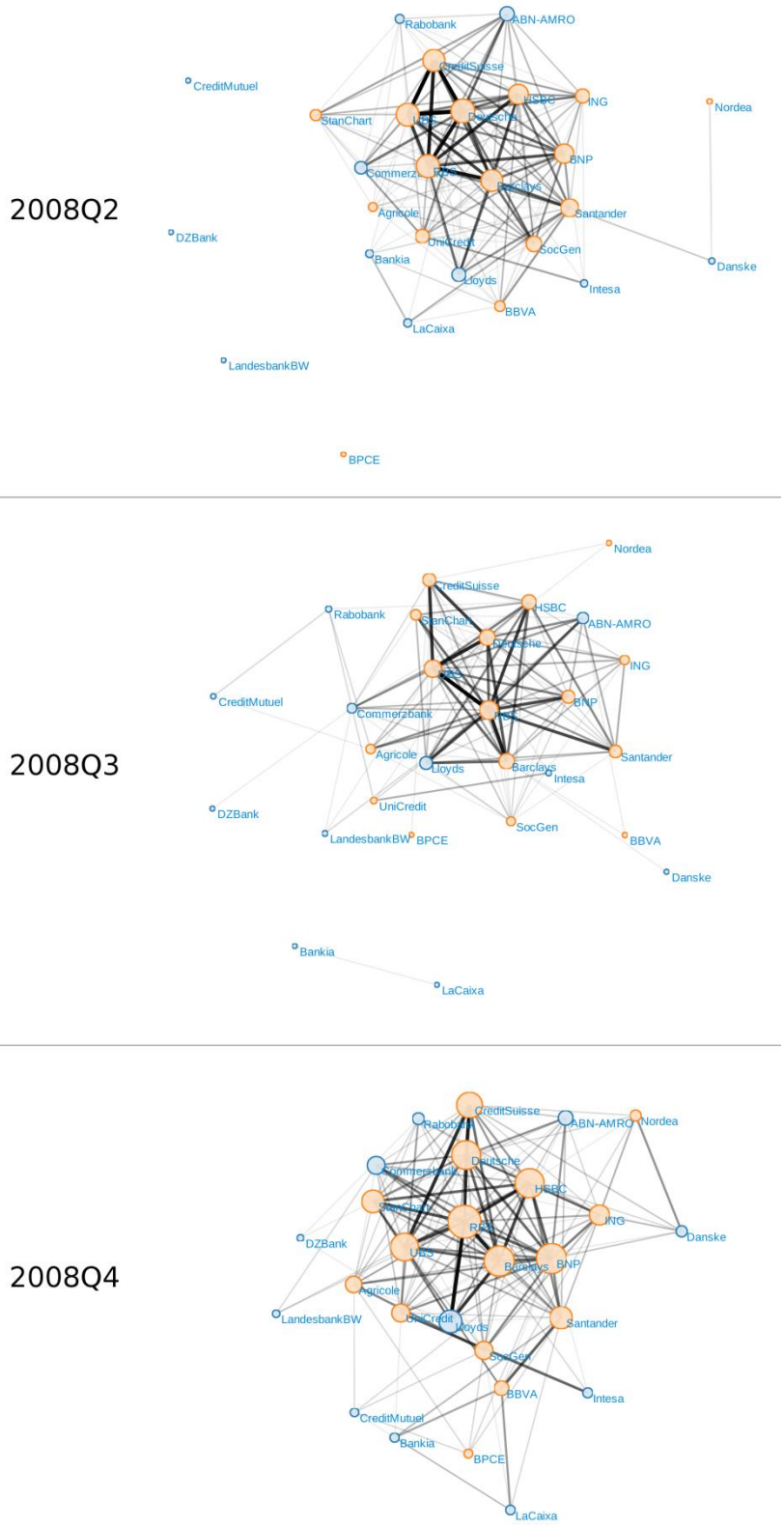


Figure 6. Network visualization for 2008Q2-Q4, each showing current link strengths and topology. Node size is relative to information centrality (α=0.1) and orange color denotes globally systemically important banks (bank labels are described in Appendix A).

3.2 Visual analysis

As a complement to the discussion on quantitative analysis of the co-occurrence networks, we briefly consider the role of visual network analysis. Our information centrality measurements highlight an interesting pattern in 2008Q2-Q4 that we inspect further visually. The second and fourth quarters have relatively high global information centrality, whereas there is a temporary dip in the third quarter. The networks in Figure 6 show visualized snapshots of each quarter, where the changes in patterns of connectivity can be studied in more detail. It shows a sparser topology for Q3 than in both Q2 and Q4, as reflected by the measurement. In addition, the visualization allows for studying local patterns, e.g., how the connection between the two Scandinavian banks Nordea and Danske Bank (right side of figure) changes.

Even though visual inspection can provide valuable insight, it may be hard to reliably and precisely compare changes in specific aspects, such as centrality of single nodes or centralization of the whole network, based on the network visualization. This underlines the importance of backing visual analysis with quantitative measures, such as encoding node size with information centrality. Still, the combination of both approaches is posed to provide the best possibilities for understanding the properties of the network, through a mixed process of exploration and focused inspection. Plots of quantitative measures and network visualizations for exploration can be presented separately, or the presentation of these data may be combined. The visual representations in Figure 6 represent information centrality as node size, which in combination with the force-directed node positioning provides support for visually assessing node centrality in more general terms.

4 DETERMINANTS OF INFORMATION CENTRALITY

Analysis thus far attempted to convince that information centrality captures the notion of system-wide importance of a bank in terms of financial discourse. Yet, little was done to provide a deeper interpretation of what information centrality signifies. This section explores potential determinants of information centrality. We explain centrality with a large number of bank-specific risk drivers, as well as country-specific macro-financial and banking sector variables, beyond controls for bank size. Further, we also assess the extent to which information centrality explains banks' risk to go bad, and compare it to more standard measures of size.

4.1 Data

We complement the textual data, and therefrom derived centrality measures, with bank-level data from financial statements and banking-sector and macro-financial indicators at the country level. This gives us a dataset of 24 risk indicators, spanning 2000Q1 to 2014Q1 for 27 banks, as well as distress events based upon bankruptcies and other types

of direct failures, government aid and distressed mergers. We use the distress events, as defined in Betz et al. [9].

To measure risk drivers, we make use of CAMELS variables (where the letters refer to Capital adequacy, Asset quality, Management quality, Earnings, Liquidity, and Sensitivity to Market Risk). The Uniform Financial Rating System, informally known as the CAMEL ratings system, was introduced by the US regulators in 1979. Since 1996, the rating system was complemented with Sensitivity to Market Risk, to be called CAMELS. The literature on individual bank failures draws heavily on the risk drivers put forward by the CAMELS framework. Further, we complement bank-level data with country-level indicators of risk. One set of variables describes the banking sector as an aggregate, whereas another explains macro-financial vulnerabilities in European countries, such as indicators from the scorecard of the Macroeconomic Imbalance Procedure. All bank-specific data are retrieved from Bloomberg, whereas country-level data comes mainly from Eurostat and ECB MFI Statistics.

4.2 What explains information centrality?

The essential question we ask herein is whether more central banks perform or behave differently. Following Bertay et al. [8], who assess whether and to what extent performance, strategy and market discipline depend on standard bank size measures, we conduct experiments in order to better understand what signifies information centrality. In contrast to their study, we control for more standard measures of bank size, in order to capture particular effects of information centrality. Using the above described data, we make use of standard, linear least squares regression models to conduct the following experiments (cf. Table 1):

1. Explain information centrality (IC) with bank size variables (Model 1).
2. Explain IC with CAMELS variable groups one-by-one, controlling for bank size (Models 2-7).
3. Explain IC with all CAMELS variables, controlling for bank size (Model 8).
4. Explain IC with CAMELS and country-specific variables, controlling for bank size (Model 9).

Our experiments show a number of patterns about drivers of information centrality. Table 1 summarizes all regression estimates. First, we show that size measures of total assets and total deposits statistically significantly explain information centrality, both when included individually and together in regressions. At a 0.1% level, we can show that these size variables relate to centrality, which is in accordance with the nature and aim of the measure.

Second, we also add variable groups from the CAMELS framework to assess which risk factors explain information centrality. When testing groups one-by-one, we find that loan loss provisions to total loans, the cost-to-income ratio, interest expenses to liabilities and deposits to funding are statistically significant at the 5% level and reserves to impaired assets and share of trading income at the 10% level. Large cost-to-income ratios are expected to reduce individual bank risk, whereas loan loss provisions are expected to increase risk. Yet, the estimates of the liquidity variables - interest expenses to total liabilities and deposits to funding - indicate less risk, as more deposits

is expected to be negatively and more interest expenses positively related to bank risk. The relationships of loan loss reserves and share of trading income are potentially ambiguous, as higher reserves should correspond to a higher cover for expected losses but could also proxy for higher expected losses and trading income might be related to a riskier business model as a volatile source of earnings but investment securities are also liquid, allowing to minimize potential fire sale losses.

Third, when including all size and CAMELS variables, we still find the same variables to be statistically significant, except for the cost-to-income ratio and the share of trading income. When assessing the size variables, assets is consistently a significant predictor, whereas deposits turns insignificant in regressions that also include deposits to funding, which is likely to be a result of multicollinearity. Further, the effects of individual risk indicators are unchanged when excluding all bank size variables. Fourth, we complement the bank-specific model with country-level data by also explaining centrality with banking sector and macro-financial variables. Even though this leads to an improvement of R^2 by one third, this leaves effects unchanged, with the exception of asset quality variables. Out of the country-specific variables, statistically significant predictors are mortgages to loans, loans to deposits, real GDP growth, stock and house price growth, and the international investment position to GDP.

4.3 Information centrality as a risk driver

In the above experiments, we have showed that information centrality is partly driven by CAMELS variables, which generally represent different dimensions of individual bank risk. This does not, however, necessarily imply that information centrality is a measure of vulnerability. The next question is whether and to what extent information centrality signals vulnerable banks, particularly when controlling for CAMELS variables.

As we have distress events for the banks, and the above used risk indicators, we can easily test the extent to which information centrality aids in identifying vulnerable banks. By focusing on vulnerable rather than distressed banks, we are interested in periods that precede distress events (e.g., 24 months). In this case, we make use of standard logistic regression to attain a predicted probability for each bank to be vulnerable. This probability is turned into a binary point forecast by specifying a threshold above which we signal vulnerability. This threshold is chosen to minimize a policymaker's loss function, who has relative preferences between false alarms and missed crises. Also, we provide a so-called Usefulness measure that captures the performance of the model in comparison to not having a model (best guess of a policymaker). We assume in the benchmark case the policymaker to be more concerned about missing a crisis than giving a false alarm, which is particularly feasible for internal signals.

To test to what extent information centrality signals vulnerabilities, and how it relates to bank size variables, we regress pre-distress events. Hence, as in a standard early-warning setting for banks, we explain periods 24 months prior to distress with logistic regression. Starting out from bank importance variables, we can see in Table 2 (Models

1-4) that while none of the variables yield highly valuable predictions, assets and deposits provide more Usefulness than information centrality. The same holds also for statistical significance. Even though the bank size variables were above shown to explain information centrality, we can observe a difference in their relation to risk.

Table 1. Regression estimates on determinants of information centrality

	Estimates	(1) Size	(2) C	(3) A	(4) M	(5) E	(6) L	(7) S	(8) CAMELS	(9) No size	(10) Bank + country
	Intercept	6.21 ***	6.24 ***	6.38 ***	6.23 ***	6.24 ***	6.56 ***	6.22 ***	6.51 ***	6.56 ***	6.53 ***
	Total assets	1.07 ***	1.23 ***	1.25 ***	1.13 ***	1.00 ***	1.96 ***	1.07 ***	2.32 ***		1.84 ***
I[a]	Total deposits	1.13 ***	0.98 ***	0.82 **	1.06 ***	1.20 ***	-0.17	1.11 ***	-0.65		-0.31
C[a]	Total equity to total assets		0.22						0.02	-0.26	-0.03
A[a]	Reserves for loan losses to impaired assets			-0.26 .					-0.96 *	-1.23 **	-0.56
	Loan loss provisions to total loans			0.64 ***					0.55 **	0.85 ***	0.16
M[a]	Operating costs to operating income				-0.29 *				-0.29	-0.20	-0.52
E[a]	Return on assets					-0.20			-0.02	0.33	0.08
	Return on equity					0.14			-0.15	-0.24	0.07
	Interest expenses to total liabilities						-0.46 **		-0.34 *	-0.84 ***	-0.60 ***
L[a]	Deposits to funding						0.86 ***		0.98 ***	0.52 *	0.67 *
	Net-short term borrowing						-0.18		-0.13	0.07	0.20
S[a]	Share of trading income							0.25 .	-0.06	0.11	-0.08
	Total assets to GDP										0.47
	Non-core liability growth										-0.05
	Debt to equity										-0.22
	Debt securities to liabilities										-0.23
	Mortgages to loans										0.63 .
	Loans to deposits										-0.86 *
	Real GDP growth										-0.54 **
	Inflation										0.14
	Stock price growth										-0.76 ***
	House price growth										-0.41 .
	Long-term government bond yield										-0.28
	International investment position to GDP										0.77 .
	Government debt to GDP										0.31
	Private sector credit flow to GDP										0.17
	Multiple R^2	0.28	0.28	0.31	0.29	0.29	0.30	0.29	0.32	0.32	0.43
	Adjusted R^2	0.28	0.28	0.30	0.28	0.28	0.30	0.28	0.31	0.31	0.40

Notes: For standardized coefficients, the explanatory variables have been transformed to have zero mean and unit variance.
Signif. codes: '***', 0.001; '**', 0.01; '*', 0.05; '.', 0.10

[a] The letters of CAMELS refer to Capital adequacy, Asset quality, Management, Earnings, Liquidity and Sensitivity to market risk, and I refers to bank size.

Large banks in terms of assets are found to be more vulnerable to distress, whereas large banks in terms of deposits are found to be less so. This is likely to proxy for the business model or activities of a bank, which might be less risky when the focus is on depository functions. Moreover, deposits can be seen as a more stable funding source than interbank market or securities funding. This points to information centrality being a more general measure of interconnectedness, rather than one defined by the underlying focus of the business model. Further, when we add all CAMELS variables to the three importance measures (Models 5--8), both Usefulness and statistical significance points to better explanatory power of assets and deposits. Comparing to models with only bank importance variables, this moves Usefulness from $U_r(\mu=0.9)$ = 35% at its maximum to 63% for information centrality and 70% for assets and deposits. Likewise, when adding all country-specific variables (Models 9--12), we can still observe that the explanatory power of assets and deposits is higher than that for information centrality. At this stage, we have early-warning models that capture most of the available Usefulness, by showing a $U_r(\mu=0.9) \geq 90\%$.

The implication of the two conducted experiments jointly is that information centrality is highly correlated with bank size, both when measured in total assets and deposits, but not a measure of vulnerability. This indicates that the measure is not biased by business activities or models, which might be a factor impacting the vulnerability of a bank. Rather, we are capturing more broadly importance of a bank in terms of information connectivity in financial discourse. This property, while due to its broad nature may be a disadvantage, provides ample means for measuring interconnectedness and centrality from a wider perspective. It is worth remembering that these text-based networks are not an ending point, but rather provide a basis for more specific queries in textual sources, which might be chosen to narrow down the context of interdependence.

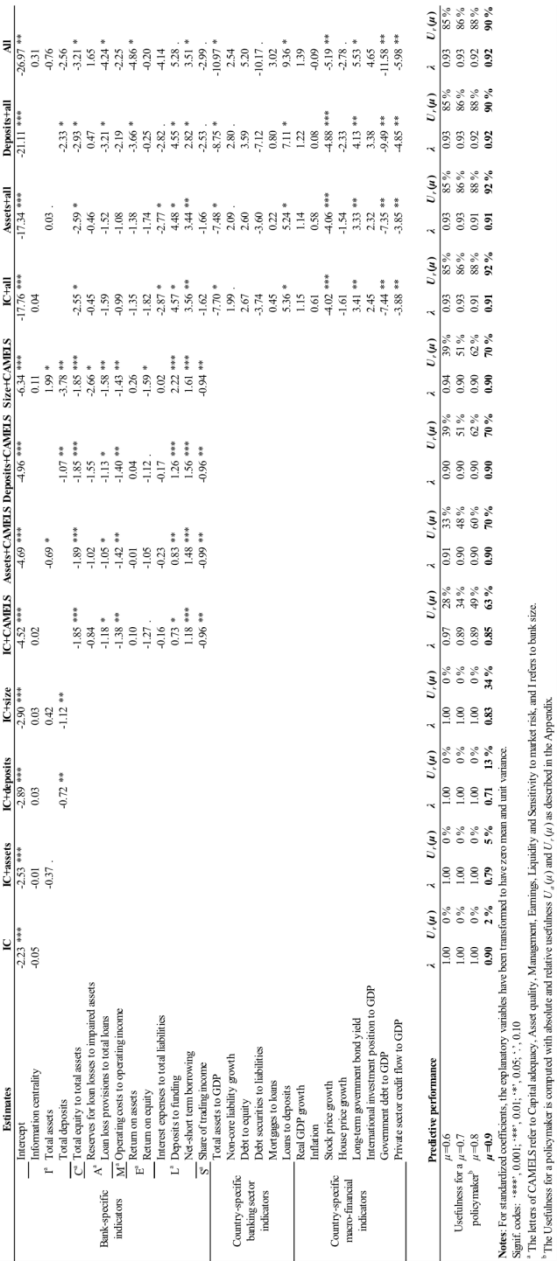
5 CONCLUSIONS

The ongoing global financial crisis has brought interdependencies among banks into focus in trying to assess systemic risk. This paper has demonstrated the use of computational analysis of financial discussion, as a source for information on bank interrelations. The approach may serve as a complement to more established ways of quantifying connectedness and dependence among banks. We have presented a text-to-network process, which has its basis in co-occurrences of bank names and can be analyzed quantitatively and visualized. To support quantification of bank importance, we proposed an information centrality measure to rank and assess trends of bank centrality in discussion. Rather than a common shortest-path based centrality measure, information centrality captures effects that might propagate aimlessly by accounting for parallel paths. Moreover, we proposed a method to stabilize the centrality measurement by applying Laplace smoothing to the link weights before calculating information centrality. To support a qualitative assessment of the bank networks, we put forward a visual, interactive interface for better illustrating network structures. This concerned not only an interface to network models, but also an interactive dashboard to better communicate quantitative network measures. Our text-based approach was illustrated on European Large and Complex Banking Groups (LCBGs) during the ongoing financial crisis by quantifying bank interrelations from discussion in 1.3M news articles,

spanning the years 2007 to 2013. However, the limitations of the current network and the underlying data occasionally lead to hazy patterns difficult to interpret and draw clear conclusions from. We suggest a number of ways these issues could be addressed in future research.

One advantage of using text data is the potentially rich thematic information it holds, which can be used to better explain or narrow the relations extracted, thereby facilitating interpretation of the network and the measures applied on top. The disadvantage of applying such filtering is that it vastly increases the data size requirements, quickly reducing a big data set into a rather scarce one. In order to apply thematic filtering to co-

Table 2. Early-warning models with information centrality



	Estimates	(1) IC		(2) IC+assets		(3) IC+deposits		(4) IC+size		(5) IC+CAMELS		(6) Assets+CAMELS		(7) Deposits+CAMELS		(8) Size+CAMELS		(9) IC+all		(10) Assets+all		(11) Deposits+all		(12) All
	Intercept	-2.23 ***		-2.53 ***		-2.89 ***		-2.90 ***		-4.52 ***		-4.69 ***		-4.96 ***		-6.34 ***		-17.76 ***		-17.34 ***		-21.11 ***		-26.97 **
	Information centrality	-0.05		-0.01		0.03		0.03		0.02						0.11		0.04						0.31
	Γ^a Total assets			-0.37 .				0.42				-0.69 *				1.99 *				0.03 .				-0.76
	Total deposits					-0.72 **		-1.12 **						-1.07 **		-3.78 **						-2.33 *		-2.56
	C^a Total equity to total assets									-1.85 ***		-1.89 ***		-1.85 ***		-1.85 ***		-2.55 *		-2.59 *		-2.93 *		-3.21 *
Bank-specific	A^a Reserves for loan losses to impaired assets									-0.84		-1.02		-1.55		-2.66 *		-0.45		-0.46		0.47		1.65
indicators	M^a Loan loss provisions to total loans									-1.18 *		-1.05 *		-1.13 *		-1.58 **		-1.59		-1.52		-3.21 *		-4.24 *
	E^a Operating costs to operating income									-1.38 **		-1.42 **		-1.40 **		-1.43 **		-0.99		-1.08		-2.19		-2.25
	Return on assets									0.10		-0.01		0.04		0.26		-1.35		-1.38		-3.66 *		-4.86 *
	Return on equity									-1.27 .		-1.05		-1.12 .		-1.59 *		-1.82		-1.74		-0.25		-0.20
	Interest expenses to total liabilities									-0.16		-0.23		-0.17		0.02		-2.87 *		-2.77 *		-2.82 .		-4.14
	L^a Deposits to funding									0.73 *		0.83 **		1.26 ***		2.22 ***		4.57 *		4.48 *		4.55 *		5.28 .
	Net-short term borrowing									1.18 ***		1.48 ***		1.56 ***		1.61 ***		3.56 **		3.44 **		2.82 *		3.51 *
	S^a Share of trading income									-0.96 **		-0.99 **		-0.96 **		-0.94 **		-1.62		-1.66		-2.53 .		-2.99 .
	Total assets to GDP																	-7.70 *		-7.48 *		-8.75 *		-10.97 *
Country-specific	Non-core liability growth																	1.99 .		2.09 .		2.80 .		2.54
banking sector	Debt to equity																	2.67		2.60		3.59		5.20
indicators	Debt securities to liabilities																	-3.74		-3.60		-7.12		-10.17 .
	Mortgages to loans																	0.45		0.22		0.80		3.02
	Loans to deposits																	5.36 *		5.24 *		7.11 *		9.36 *
	Real GDP growth																	1.15		1.14		1.22		1.39
	Inflation																	0.61		0.58		0.08		-0.09
Country-specific	Stock price growth																	-4.02 ***		-4.06 ***		-4.88 ***		-5.19 **
macro-financial	House price growth																	-1.61		-1.54		-2.33		-2.78 .
indicators	Long-term government bond yield																	3.41 **		3.33 **		4.13 **		5.53 *
	International investment position to GDP																	2.45		2.32		3.38		4.65
	Government debt to GDP																	-7.44 **		-7.35 **		-9.49 **		-11.58 **
	Private sector credit flow to GDP																	-3.88 **		-3.85 **		-4.85 **		-5.98 **
	Predictive performance	λ	$U_r(\mu)$	λ	$U_r(\mu)$	λ	$U_r(\mu)$	λ	$U_r(\mu)$	λ	$U_r(\mu)$	λ	$U_r(\mu)$	λ	$U_r(\mu)$	λ	$U_r(\mu)$	λ	$U_r(\mu)$	λ	$U_r(\mu)$	λ	$U_r(\mu)$	λ	$U_r(\mu)$
	μ=0.6	1.00	0 %	1.00	0 %	1.00	0 %	1.00	0 %	0.97	28 %	0.91	33 %	0.90	39 %	0.94	39 %	0.93	85 %	0.93	85 %	0.93	85 %	0.93	85 %
Usefulness for a	μ=0.7	1.00	0 %	1.00	0 %	1.00	0 %	1.00	0 %	0.89	34 %	0.90	48 %	0.90	51 %	0.90	51 %	0.93	86 %	0.93	86 %	0.93	86 %	0.93	86 %
policymaker[b]	μ=0.8	1.00	0 %	1.00	0 %	1.00	0 %	1.00	0 %	0.89	49 %	0.90	60 %	0.90	62 %	0.90	62 %	0.91	88 %	0.91	88 %	0.92	88 %	0.92	88 %
	μ=0.9	**0.90**	**2 %**	**0.79**	**5 %**	**0.71**	**13 %**	**0.83**	**34 %**	**0.85**	**63 %**	**0.90**	**70 %**	**0.90**	**70 %**	**0.90**	**70 %**	**0.91**	**92 %**	**0.91**	**92 %**	**0.92**	**90 %**	**0.92**	**90 %**

Notes: For standardized coefficients, the explanatory variables have been transformed to have zero mean and unit variance.

Signif. codes: '***': 0.001; '**': 0.01; '*': 0.05; '.': 0.10

[a] The letters of CAMELS refer to Capital adequacy, Asset quality, Management, Earnings, Liquidity and Sensitivity to market risk, and I refers to bank size.

[b] The Usefulness for a policymaker is computed with absolute and relative usefulness $U_a(\mu)$ and $U_r(\mu)$ as described in the Appendix.

occurrence links between banks, we recommend more sophisticated semantic analysis to increase recall. For instance, distributional semantic methods [28] could be used to extend a set of seed keywords, or probabilistic topic modeling [10] could be applied to the corpus to identify topics of interest and the related subset of articles. Furthermore, combining sentiment analysis with our bank relation extraction could constitute another interesting way to distinguish the nature of mapped relations. Sentiment analysis has been applied to classify company-related information from financial news in regards to the effect on their stock price (e.g., [23]), an approach that could hold considerable potential in the area of systemic risk analysis as well.

REFERENCES

[1] Viral. Acharya, Lasse Pedersen, Thomas Philippon, and Matthew Richardson. Measuring systemic risk. 2012.

[2] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. Science, 286(5439):509–512, 1999.

[3] M. Bostock, V. Ogievetsky., J. Heer. D3: Data- driven documents. IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis) (2011).

[4] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. Proceedings of the National Academy of Sciences of the United States of America, 101(11):3747–3752, 2004.

[5] S. Battiston, M. Puliga, R. Kaushik, P. Tasca, and G. Caldarelli. Debtrank: Too central to fail? financial networks, the fed and systemic risk. Scientific Reports, 2:541, 2012.

[6] A. C. Bertay, A. Demirgüc-Kunt and H. Huizinga. Do we need big banks? evidence on performance, strategy and market discipline. Journal of Financial Intermediation, 22(4):532– 558, 2013.

[7] F. Betz, S. Oprica, T. Peltonen, and P. Sarlin. Predicting distress in European banks. Journal of Banking & Finance, forthcoming.

[8] David M Blei. Probabilistic topic models. Communications of the ACM, 55(4):77–84, 2012.

[9] Stephen P Borgatti. Centrality and network flow. Social networks, 27(1):55–71, 2005.

[10] Claudio EV Borio and Mathias Drehmann. Towards an operational framework for financial stability:" fuzzy" measurement and its consequences. Number 284. Bank for International Settlements, Monetary and Economic Department, 2009.

[11] Ulrik Brandes and Daniel Fleischer. Centrality measures based on current flow. STACS 2005, pages 533–544, 2005.

[12] Eugenio Cerutti, Stijn Claessens, and Patrick McGuire. Systemic risk in global banking: what can available data tell us and what more data are needed? 2012.

[13] Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. Computer Speech & Language, 13(4):359–393, 1999.

[14] Giovanni Dell'Ariccia, Enrica Detragiache, and Raghuram Rajan. The real effect of banking crises. Journal of Financial Intermediation, 17(1):89–112, 2008.

[15] Vasant Dhar. Data science and prediction. Communications of the ACM, 56(12):64–73, 2013.

[16] Edsger W Dijkstra. A note on two problems in connexion with graphs. Numerische mathematik, 1(1):269–271, 1959.
[17] European Central Bank. Financial Stability Review, November 2013.
[18] Financial Stability Board. 2013 update of group of global systemically important banks (G- SIBs), November 11 2013.
[19] Nikolaus Hautsch, Julia Schaumburg, and Melanie Schienle. Financial network systemic risk contributions. Technical report, CFS Working Paper, 2013.
[20] Matthew O Jackson and Brian W Rogers. Meeting strangers and friends of friends: How random are social networks? The American economic review, pages 890–915, 2007.
[21] Luc Laeven and Fabian Valencia. Resolution of banking crises: The good, the bad, and the ugly. International Monetary Fund, 2010.
21[22] Burton G Malkiel. The efficient market hypothesis and its critics. Journal of economic perspectives, pages 59–82, 2003.
[23] Pekka Malo, Ankur Sinha, Pyry Takala, Oskar Ahlgren, and Iivari Lappalainen. Learning the roles of directional expressions and domain concepts in financial news analysis. In Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on, pages 945–954. IEEE, 2013.
[24] Mark EJ Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. Physical review E, 64(1):016132, 2001.
[25] Tore Opsahl, Filip Agneessens, and John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks, 32(3):245–251, 2010.
[26] Arzucan Ozgür, Burak Cetin, and Haluk Bingol. Co-occurrence network of reuters news. International Journal of Modern Physics C, 19(05):689–702, 2008.
[27] P. Sarlin. On policymakers' loss functions and the evaluation of early warning systems. Economics Letters, 119(1):1–7, 2013.
[28] Karen Stephenson and Marvin Zelen. Rethinking centrality: Methods and examples. Social Networks, 11(1):1 – 37, 1989.
[29] Peter D Turney, Patrick Pantel, et al. From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, 37(1):141–188, 2010.
[30] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. Nature, 393(6684):440–442, 1998.
[31] Jonathan D Wren, Raffi Bekeredjian, Jelena A Stewart, Ralph V Shohet, and Harold R Garner. Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics, 20(3):389–398, 2004.

APPENDIX A: DATA

Table 3. A list of banks and their labels.

Label	Name	Label	Name
Agricole	Credit Agricole Groupe	ABN-AMRO	ABN AMRO Bank NV
BBVA	Banco Bilbao Vizcaya Argenta	Bankia	Bankia SA
BPCE	Groupe BPCE	Commerzbank	Commerzbank AG
BNP	BNP Paribas	CreditMutuel	Credit Mutuel Group
Barclays	Barclays PLC	DZBank	DZ Bank AG
CreditSuisse	Credit Suisse Group AG	Danske	Danske Bank A/S
Deutsche	Deutsche Bank AG	Intesa	Intesa Sanpaolo
HSBC	HSBC Holdings PLC	LaCaixa	La Caixa
ING	ING Bank NV	LandesbankBW	Landesbank Baden-Württemberg
Nordea	Nordea Bank AB	Lloyds	Lloyds Banking Group PLC
RBS	Royal Bank of Scotland	Rabobank	Rabobank Group
Santander	Banco Santander SA		
SocGen	Group Societe Generale SA		
StanChart	Standard Chartered PLC		
UBS	UBS AG		