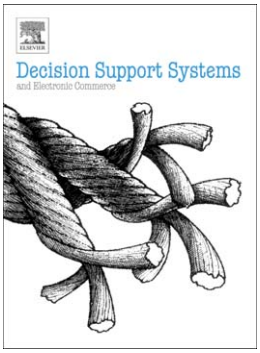


Bankruptcy prediction for SMEs using relational data

Ellen Tobback, Tony Bellotti, Julie Moeyersoms, Marija Stankova, David Martens

PII: S0167-9236(17)30138-0
DOI: doi:10.1016/j.dss.2017.07.004
Reference: DECSUP 12865

To appear in: *Decision Support Systems*

Received date: 5 January 2017
Revised date: 19 May 2017
Accepted date: 14 July 2017

Please cite this article as: Ellen Tobback, Tony Bellotti, Julie Moeyersoms, Marija Stankova, David Martens, Bankruptcy prediction for SMEs using relational data, *Decision Support Systems* (2017), doi:10.1016/j.dss.2017.07.004

Bankruptcy prediction for SMEs using relational data

Ellen Tobback[a,*], Tony Bellotti[b], Julie Moeyersoms[a], Marija Stankova[a], David Martens[a]

[a] *Department of Engineering Management, University of Antwerp, Belgium*
[b] *Department of Mathematics, Imperial College London, UK*

Abstract

Bankruptcy prediction has been a popular and challenging research area for decades. Most prediction models are built using financial figures, stock market data and firm specific variables. We complement such traditional *low-dimensional* data with *high-dimensional* data on the company's directors and managers in the prediction models. This information is used to build a network between small and medium-sized enterprises (SMEs), where two companies are related if they share a director or high-level manager. A smoothed version of the weighted-vote relational neighbour classifier is applied on the network and transforms the relationships between companies into bankruptcy prediction scores, thereby assuming that a company is more likely to file for bankruptcy if one of the related companies in its network has already failed. An ensemble model is built that combines the relational model's output scores with structured data and is applied on two data sets of Belgian and UK SMEs. We find that the relational model gives improved predictions over a simple financial model when detecting the riskiest firms. The largest performance increase is found when the relational and financial data are combined, confirming the complementary nature of both data types.

Keywords: Data mining; Relational data; Network analysis; Bankruptcy prediction; SME

*Corresponding author
Email address: ellen.tobback@uantwerpen.be (Ellen Tobback)

1. Introduction

Bankruptcy prediction is a widely studied topic due to its importance for the banking sector. The current volume of outstanding debt to non-financial firms in Belgium is about 122 billion euros, which is 123% of GDP as measured in the first quarter of 2015 [26]. The size of corporate lending makes sound lending decisions a matter of national interest. To counter the adverse effects of these high exposures, Basel II and III have introduced capital requirements that are more sensitive to risk. For many SMEs this implies that banks are charging a higher risk premium [5]. Investing in improved bankruptcy prediction models is therefore in the interest of both the banks and the clients, as better predictions will reduce risk and lower the subsequent risk premia.

Research on bankruptcy prediction has largely focused on traditional data such as financial ratios, stock data or macroeconomic data [10, 34]. However, it is often noted that the (in)competence of the managerial team has a great influence on a company's chance of survival [29]. To measure a business manager's or board member's competence, one could take a look at the business history of this person. When a person was involved in a bankruptcy case in the past, banks will be reluctant to grant this person a loan for the start-up of a new firm. Notwithstanding the clear importance of the management's competence and historical success/failure, most research on bankruptcy prediction does not take this kind of data into account. In this paper, we intend to fill this research gap and try to predict bankruptcy using both traditional, financial data and fine-grained data on person-related relationships.

We exported data from Belfirst and Fame [1], databases containing financial reports and statistics on respectively Belgian and UK companies. This data can be categorized into traditional data and relational data. The traditional, dense data are mostly financial ratios such as the current ratio, debt ratio and return on assets; and firm-specific data such as

[1] Both databases are managed by Bureau van Dijk.

the company's age and sector. The relational data captures the links between companies at the board and management level. Using relational data in a bankruptcy setting we start from two possibly related assumptions: (i) if a company is linked to many bankrupt firms, it will have a higher probability of becoming bankrupt and (ii) the management has an influence on the performance of the company and incompetent or fraudulent managers can lead a company into bankruptcy. The latter assumption has already been demonstrably investigated [6, 29], the former assumption is a possible derivation from the latter, however we do not exclude the fact that there could be other causes leading to the first assumption, such as a loss of supply or demand if both companies were trade partners.

The contributions to the literature are five-fold. Empirical research on corporate bankruptcy focuses mainly on innovations in modelling techniques and only to a lesser extent on innovations in the feature space. To the best of our knowledge, we are the first to use fine-grained data about relationships between companies for credit scoring. Secondly, whereas most studies focus on a sample of companies from a certain country, we used *all* Belgian and UK SMEs that publish financial statements leading to two data sets of around 400,000 and 2,000,000 companies. Thirdly, we use a completely out-of-time set-up and take into account that healthy companies in the training set can become bankrupt companies in the test set and apply a tailored 'leave-many-out' procedure to deal with the double occurrences. Fourthly, we indicate how much data one should collect, by investigating if resigned directors and old bankruptcies have an influence on the prediction performance. Finally, we present a smoothed version of the weighted vote relational neighbour (wvRN) classifier and show that it increases the lift of the riskiest firms.

The remainder of this paper is organized as follows. Section 2 defines the research question and provides the reader with a concise overview of the relevant literature and progress. Section 3 details the transformation from relational data into an SME network. Section 4 describes the financial performance indicators and relational data used in this study. Section 5.1 provides a detailed description of our methodology and Section 5.2

summarizes and analyses the empirical results. Finally, Section 6 discusses the conclusions and provides insights for future research.

2. Literature overview

2.1. Definition and terminology

In this study the term bankruptcy is used interchangeably with failure and default, where the notion of bankruptcy refers to the legal status of an entity when it cannot repay its owed debts. We will test our proposed methodology on the SME network of two countries: Belgium and the UK. Both countries have a different bankruptcy law. In Belgium, bankruptcy is part of the commercial law, which implies that only merchants can go bankrupt [13]. According to Article 2 in bankruptcy law, the directors of a commercial company that has durably ceased making payments or that has lost its creditworthiness, are legally obligated to petition for the company's bankruptcy. In the UK, the term bankruptcy is used as a legal term for individuals only. The appropriate term for companies that cannot repay their debt is *insolvency*. The UK insolvency law prescribes four different procedures in case of insolvency [17]. In the first three procedures the primary goal is an attempt to rescue the firm. The fourth procedure is the liquidation of the firm, where a liquidator is appointed and the firm's assets are sold. This procedure can be compared to the bankruptcy settlement in Belgian law. For consistency when discussing bankruptcy of both Belgian and UK firms, we say a firm is bankrupt when it has been liquidated due to insolvency. Note, however, that this is contrary to the legal UK definition.

2.2. Data mining and bankruptcy prediction

There is a vast amount of research on bankruptcy prediction, going all the way back to the 1960's. The earliest research applied a univariate approach, comparing one historical ratio at a time [8]. The multivariate approach to bankruptcy prediction was first introduced by Altman [2] and Ohlson [27]. The former used multivariate discriminant analysis to find a linear function that distinguishes between healthy and bankrupt firms resulting in the

famous Z-score [2], while the latter used logistic regression to estimate the probability of bankruptcy for each firm [27]. Both added financial ratios as inputs to their prediction models.

Since the 1990's the focus has shifted towards artificially intelligent expert models, such as neural networks and Support Vector Machines (SVM). Multilayer neural networks are reported to significantly outperform both logistic regression [40] and Multivariate Data Analysis (MDA) [39] and a number of studies have successfully applied SVM for corporate bankruptcy prediction [32] and shown that they are competitive with MDA [25] and logistic regression [37, 25]. The performance improvement of bankruptcy prediction with these intelligent techniques indicates that the influence of financial ratios on a firm's health has non-linear properties. However the choice of non-linear, black-box models decreases the comprehensibility of the bankruptcy predictions. Hence, in a practical setting discriminant analysis and logit models remain dominant.

Empirical research on corporate bankruptcy focuses mainly on innovations in modelling techniques and only to a lesser extent on innovations in the features space. The most frequently used features are firm or industry specific information and performance indicators. Amongst the performance indicators, the current ratio and the Return on Assets ratio are the most commonly used factors [10]. More recent studies have investigated the predictive power of market/stock data and macroeconomic variables [34]. What all the aforementioned studies have in common, is that they focus on dense, structured data (mainly financial ratios). Recent advances in the use of sparse fine-grained data[2] have shown that they add incremental predictive power to the models. Hence, this kind of data can be useful in bankruptcy predictions as well. Various studies highlighted the management's role in the failure process. Ooghe and De Prijcker [29] distinguish three types of shortcomings that may cause corporate bankruptcy: (i) a lack of competences and skills (ii) insufficient motivation and (iii) certain personal characteristics such as risk affinity, over-optimism and

[2]See e.g. [38] for behavioural data and [18] for network data.

haste. Ma et al. [21] explore the importance of management capability on SME performance and, based on transactional data, show that it is valuable to include management capability as a factor to model credit risk. Baldwin et al. [6] found that managerial weakness was the main cause of small business bankruptcy in Canada. We can account for these influences on bankruptcy by adding sparse relational data as features to our prediction models.

In this study, we focus on SMEs and exclude large firms from our data set. Altman and Sabato [3] found that creating a specific model for SMEs leads to significantly more prediction power than using a generic corporate prediction model. Because unlisted firms are often allowed to file an abbreviated financial statement, much of the information that is necessary to calculate the accounting ratios that are typically used to model failure is not available for SMEs. Altman et al. [4] therefore suggest to use non-financial information to account for the missing financial information. They include information on the type, sector, size and age of the firm as well as information on the reporting and compliance, and operational risk. They find that non-financial data significantly increases the bankruptcy prediction power of SME risk models. With this study, we add to the literature on SME model risk prediction and combine financial information with non-financial, firm specific characteristics, including information on the firm's directors and managers.

2.3. Challenges and success of relational data

Relational data is data that defines relationships between entities. The use of relational data has already proven to be successful in other domains such as targeted advertising [18] and fraud detection [19]. Two major categories of relational data can be distinguished: real network data and pseudo-network data. In a real network, two nodes are connected because a certain form of direct communication has taken place between them. In a pseudo or implied network, two nodes are connected because they have a common interest, activity or asset: e.g. they have watched the same videos [38] or paid to the same entities [23]. The network is implied as there is no evidence that both nodes have ever communicated with each other. In this research we create an implied network by linking companies based on

the shared board members/managers.

The nature of relational data requires a different approach than the traditional financial data. One of the main challenges of using network data is the transformation from its rough form, i.e. a list of managers per company, to a structured form, i.e. a weighted sparse matrix where the weights denote the strength of the link between two entities (here companies). The sparsity of the data set requires a large sample that contains all relevant neighbours for each entity in the data set. Next, specifically tailored learners have to be used to obtain a prediction score for each entity with unknown class in the network. Relational learners are powerful tools that can handle the low event rate of most network data sets. Section 5.1 explains in detail how we processed the relational data.

3. A network of SMEs

The proposed methodology starts by creating a network of SMEs. Two firms are linked if they have an entity in common. There exists a large variety of entities that can be used to connect two firms, from directors and managers to suppliers and clients. In this paper, we use information about past and current directors and managers to link companies. Hence, we create a network of SMEs where two companies are linked if they share or have shared a member of the board of directors and/or the management board. Using this data builds upon the assumption that being linked to one or more bankrupt companies increases your own probability of default. Figure 1 illustrates our methodology in line with the approach of Stankova et al. [33]. We project a bipartite graph with links between companies and their directors/managers into a weighted unigraph that links companies to each other. Relational learners are applied to the network of companies.

To each link between companies, a strength should be assigned. There are several possibilities to define the edge weight, however we limit the scope to a (weighted) aggregation of the number of managers/directors the firms share. When aggregating the top nodes, we consider top node weighting schemes, as reported in Table 1 [33]. Most of these weight functions downweigh top nodes with a high degree, assuming these provide less information.

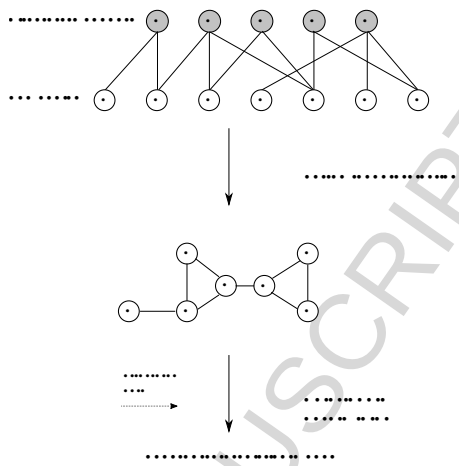


Figure 1: Create a weighted projection from the bigraph (above). The board members/managers are the top nodes and the companies are the bottom nodes. Relational learners are applied to the resulting network of companies (below). The network model is then compared to a base model containing financial data and an ensemble model.

In the SME-context, it is acceptable to assume that a manager with many positions will have less time to invest in the company and therefore less time to influence a company's performance. A special top node weight function is the *time function*. This top node weight is different for every ij-combination and represents the time (in days) person k has spent at firm i (δ_i^k) and firm j (δ_j^k) divided by the total number of days (n^k) person k has been active (in all companies of the training set).

As a relational learner, we apply the weighted-vote Relational Neighbour (wvRN) classifier [22]. It is a simple, yet powerful classifier that uses the network structure to calculate a bankruptcy probability score for a company as a weighted average of its j neighbours' probability scores (see Equation 1). The classifier is based on the property of assortativity (also known as homophily in social network theory [24]), as it makes the assumption that the connected companies are similar and therefore more likely to belong to the same class. This is aligned with our premise that the companies related through the same man-

Table 1: Top node weight functions. **s_k is the weight of node (i.e. director/manager) k, d_k is the degree of person k, d_k^c is the degree of person k for bottom nodes of class c, N is the total number of bottom nodes (i.e. firms), δ_i^k is the number of days director/manager k has spent at firm i and n^k is the total number of days person k has been active in a firm.**

Top node weight function	Formula
Inverse degree	$s_k = \frac{1}{d_k}$
Inverse frequency	$s_k = \log_{10} \frac{N}{d_k}$
Hyperbolic tangent	$s_k = \tanh(\frac{1}{d_k})$
Adamic and Adar	$s_k = \frac{1}{\log_{10}(d_k)}$
Delta function	$s_k = \frac{2}{d_k(d_k-1)}$
Class-degree ratio	$s_k = \frac{d_k^c}{d_k}$
Time function	$s_{ijk} = \frac{\delta_i^k + \delta_j^k}{n^k}$

agers/directors are likely to exhibit similar bankruptcy behaviour due to the incompetence or fraudulent intentions of the managerial team. The wvRN classifier has the following form:

$$P(L_i = c|N(i)) = \frac{1}{Z} \sum_{j \in N(i)} w_{ij} P(L_j = c|N(j))$$

$$\text{where the normalization factor Z is equal to } \sum_{j \in N(i)} w_{ij} \tag{1}$$

$$\text{and } w_{ij} \text{ is the sum of all shared directors/managers: } w_{ij} = \sum_{k \in N_T(i) \cap N_T(j)} s_k$$

Equation 1 calculates the probability that the label L of company i equals c, with c a binary indicator of bankruptcy, given its neighbours $N(i)$ in the unigraph projection. The edge weight w_{ij} between company i and its neighbour j is the sum of the top node weights s_k of all shared top nodes, where N_T denotes the top node neighbours in the bipartite graph. The resulting bankruptcy probability score is the weighted sum of the bankruptcy probabilities of a company's neighbours. In this study, the neighbour's bankruptcy probability is set to either 0 or 1, depending on whether they went bankrupt or not.

3.1. Smoothed probabilities

When estimating bankruptcy probabilities, wvRN will assign boundary values to nodes with only one neighbour, i.e. one or zero depending on whether the neighbour is bankrupt or not. This implies that historical failures will be penalized too much when using this version of wvRN. Similarly, the method will assign boundary values when the node is surrounded by neighbours of only one type. However, a firm connected to healthy firms still has a certain probability of bankruptcy. To solve this problem, we calculate a smoothed version of the probability estimate using the concept of additive smoothing. Given N observations and x_c the number of times a certain event c occurs in these observations, the smoothed probability has the following form [14]:

$$\hat{P}(c) = \frac{x_c + \alpha}{N + \alpha d} \tag{2}$$

with d the number of categories and α the smoothing parameter. We set $\alpha = 1$ for Laplace smoothing. We apply the same logic to our data set and prediction problem. The trial outcomes x_c are the weighted probabilities $w_{ij}P(L_j = c|N(j))$ and the number of trials N is the weighted denominator Z. Traditional additive smoothing starts from the prior assumption of equal probabilities $1/d$ for each class. This assumption is not valid for the two data sets used in this study, therefore we replace the uniform probability $1/d$ by the incidence rate μ_c of the training set. We obtain the following smoothed equation:

$$P(L_i = c|N(i)) = \frac{\sum_{j \in N(i)} w_{ij} P(L_j = c|N(j)) + 2\mu_c}{Z + 2} \tag{3}$$

As a result, a firm with no neighbours will be assigned the bankruptcy rate μ_1 of the training set.

4. Empirical results: SME networks for Belgium and the UK

4.1. Data

We gathered data from Belfirst and Fame on $400,000+$ Belgian SMEs and $2,000,000+$ UK SMEs, covering the time-period 2011 to 2014. Both databases are publicly available,

though access requires the payment of a subscription fee. The classification of companies as SMEs complies with the definition of the Basel II Capital Accords, where companies are granted an SME-status if the reported yearly sales for the consolidated group the firm belongs to are less than EUR 50 million [7]. The databases Belfirst and Fame add the extra requirement that the number of employees is less than 1000. We exported financial ratios, the name and unique identifier of the current and past directors and managers, the date of incorporation, the industry code, the size of the company (small or medium) and information about the state of the company (bankrupt or active).

4.2. Financial performance indicators

Table 2: Financial performance indicators

Variables	Category	Used by
Debt to total assets ratio	Leverage/Solvency	[1, 25]
Current ratio	Liquidity	[12, 31]
Cash flow to equity ratio	Profitability	[20]
Return on equity	Profitability	[9, 12, 20]
Profit/Loss	Profitability	[28]
Return on total assets	Profitability	[9, 31]

Generally, the ratios can be divided into three categories: solvency, liquidity and profitability. Insufficient solvency and liquidity as well as low profitability are all factors that can lead to bankruptcy if not resolved by management. It is therefore important to include at least one indicator from each category in the bankruptcy prediction model. Table 2 lists the financial ratios that are used in this study. We chose a selection of financial performance indicators that covers all categories (liquidity, solvency and profitability). Because SMEs in Belgium and especially in the UK are allowed to submit a reduced version of the financial statement, many of the typically used financial variables (e.g. the majority of variables proposed by Altman [2] and Ooghe and Van Wymeersch [30]) are not available for most firms in our data set. We chose variables that were available in small Belgian companies (as these

have a more restricted version of the financial statement than the medium firms). The first category, solvency, represents the company's ability to meet its long-term obligations and can be represented by the equity ratio. This ratio measures the part of total assets that is financed by investor's equity and gives an indication of the long-term debt burden of a firm. The lower the equity ratio, the more leveraged the firm is and the higher the risk of insolvency. The second category, liquidity, indicates a company's ability to meet its short-term obligations. The 'current' ratio is a good indicator of a firm's liquidity as it represents the company's ability to quickly convert assets into cash without any loss. The ideal current ratio is two-to-one, which means the current assets are double the current liabilities [30]. It is important to measure both solvency and liquidity to assess a firm's performance, as they represent a company's long-term and short-term chance of survival, respectively. The last category, profitability, measures the economic viability of a company. Over the long term, the firm must be profitable to ensure that both liquidity and solvency are maintained. Return on assets (ROA) measures the management's ability to convert its assets into profit. A higher ROA indicates that the company is able to generate more earnings with less investment. Return on Equity (ROE) measures the profit the company generates with the shareholder's equity. Finally, the cash flow to equity ratio indicates the company's capacity to create gross income, independent of the use [30]. Insufficient liquidity, insolvency and low profitability (or loss) are warning signs of a possible future bankruptcy, however, the prediction is not perfect. It might be that the company chooses not to publish its financial statement in periods of financial stress, that the financial statement is manipulated, that the company's behaviour is fraudulent or that - due to the delay in publication - the deterioration is not noted in time. Combined with the fact that mismanagement can lead to a company's failure, we supplement financial performance indicators with relational data.

4.3. Relational data

There are more than 700,000 directors/managers in the Belgian data set and more than 4.5 million in the UK data set. On average, there are 2.7 listed directors/managers in a

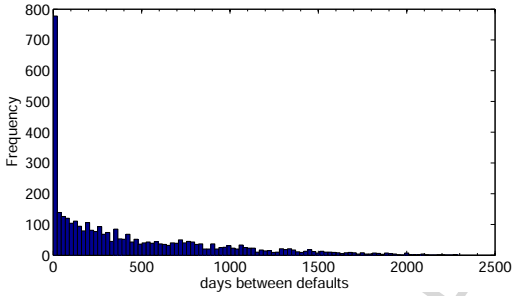


Figure 2: The number of days between the defaults of two linked firms.

Belgian SME and 4.1 in a UK SME. The histogram in Figure 2 displays the frequency of the number of days between the bankruptcy of a Belgian SME from the test set and all its linked bankrupt firms, including other firms in the test set. Note that a large number (i.e. 394) of linked firms go bankrupt on the exact same date. This information could never be used by the bank. However, every linked bankruptcy with a delay of at least one day provides information that can be employed by the bank. Short delays (e.g. up to a couple of days) will not find its influence in the bank's credit scores, which is a major limitation to the proposed methodology. However, once the bankruptcy of a linked firm is noticed, banks can take actions (such as limiting the access to a credit line) until the credit score has been re-evaluated. For 2534 linked firms, the time between bankruptcy is minimum one day and maximum one year. This information cannot be used in the model, since all these linked firms are part of the test set. Moreover, due to our ensemble set-up which is further explained in Section 5.1, we have a one-year gap between the network training set and our test set. The companies that defaulted in 2013 are not added to the network, since they are used in the ensemble model's training set. Hence, we cannot operate on the first part of the histogram. However, once the weights of the ensemble model are estimated, the need for a second training set disappears. In practice, banks will thus be able to link the companies to companies that defaulted or were active in the year prior to the prediction date. We try to predict default one year ahead. However, as Figure 2 shows, the influence

of a defaulted company on its linked companies can be delayed. The prediction scores of the relational learners can be interpreted as warning signs as well, i.e. companies with a high score are linked to many or only bankrupt firms and should be closely monitored. Figures 3 and 4 plot the degree distributions for the number of managers (and directors) that are/were part of the firm and the number of firms a manager is/has been part of. While most firms have only a limited number of managers, and most managers worked at a limited number of firms, the degree distributions show that 29% of the UK firms and 15% of the Belgian firms have at least 5 managers and that 4% of the UK managers and 2.4% of the Belgian managers have worked at more than 4 firms.

Figure 3: Degree distributions for the number of managers per firm and the number of firms per manager for the Belgian data set.

(a) Number of managers per firm (b) Number of firms per manager

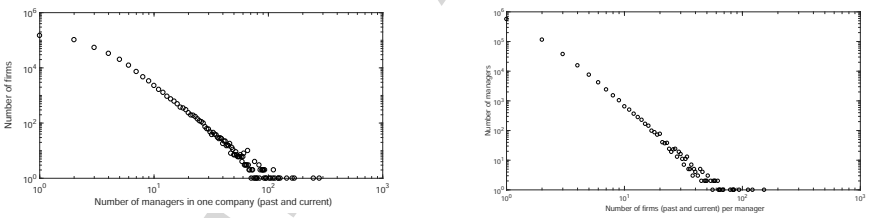
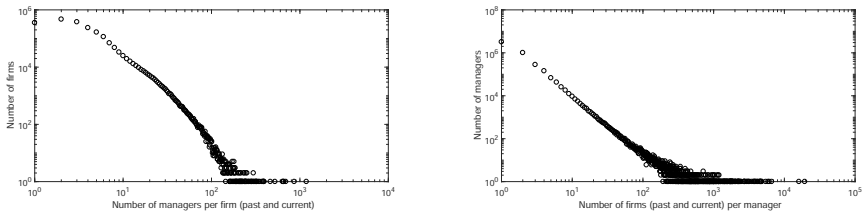


Figure 4: Degree distributions for the number of managers per firm and the number of firms per manager for the UK data set.

(a) Number of managers per firm (b) Number of firms per manager



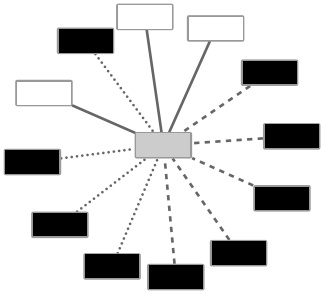


Figure 5: Example of a test node in 2014 (gray) and its connections with healthy companies (white) and companies that defaulted before 2013 (black). The different line types represent the different managers through which the test node is connected to the surrounding nodes. One manager (full lines) is responsible for the connections with the three healthy companies, the two remaining managers (striped and dotted lines) connect the company to bankrupt firms.

5. Study design and results

5.1. Study design

The goal of the study is to predict whether a company will go bankrupt in 2014 using both financial and relational data. Financial data and relational data are heterogeneous data types that have different modelling requirements, therefore we create three separate models: a financial model, a relational model and an ensemble model that adds the relational bankruptcy prediction score as extra variable next to the financial data in a linear model. As input variables to the financial model, we chose a selection of financial ratios that have been shown to be predictive of bankruptcy. All variables are normalised between -1 and +1 using min-max rescaling after the removal of outliers by the Winsorized mean procedure [35]. Due to the delay on the publication of the financial statements, we are obliged to consult the financial statement of a year prior to the observation date. This means that if we want to predict if the companies active on January 1 2014 will go bankrupt within the next year, we will have to use the financial ratios of 2012.

Some firms have missing values for all financial ratios. There are two major reasons for these missing values: (i) the firm did not publish a financial statement and (ii) the firm was

founded in 2012. It can be expected that missing values due to a missing financial statement are a predictive factor for bankruptcy. On average 10% of the values for each financial ratio in the Belgian data set are missing values, half of these can be explained by the fact that the companies are newly founded. To distinguish between newly founded and other firms, we add a dummy variable that has value 1 if the firm was founded in the respective year and 0 otherwise. Mean imputation is used: all missing values are replaced by the average value of the training set and accompanied by a 'missing values'-dummy. The number of missing values for UK data set is much larger, as the reporting requirements for UK SMEs are less complete. While 22.6% of the firms are recently founded, on average 70% of the values for each financial ratio are missing. The current ratio and the solvency ratio have the least amount of missing values: on average 30% of the values are missing. Clearly, the Belgian and UK data sets are very different, we deal with this issue by learning the financial models separately for each data set. To control for age- and industry-specific effects, we included the normalised number of years since the foundation of the company and the 21 dummy-encoded industry SIC (for the UK) or NACE (for Belgium) codes (sectors A to U). Please note that the final financial model is a simple linear model that can be much improved upon. The choice of a linear model allows easy investigation of the incremental value of the network score in the ensemble model.

Table 3 reports the relational data characteristics of the Belgian and UK data sets. Due to the ensemble set-up, it is paramount that we work with two training sets: one training set to build and train the network and one training set to build and train the ensemble model. The first training set includes all companies that were incorporated before 2013, excluding those that ceased operations for reasons other than bankruptcy. Each company is assigned a label 0 if it was still active at the end of 2012 and 1 if it had gone bankrupt. The second training set includes all companies that were active at 1st January 2013 (i.e. the observation date) and a label indicating whether they went bankrupt in the course of 2013. Finally, the test set includes all companies that were active at 1st January 2014 (i.e.

Table 3: Data characteristics of the Belgian and UK data sets. N_i is the number of firms in the respective sample, d_i the number of bankruptcies, p_i the number of directors, n_i the unique number of company-director links, An_i the average number of directors per firm and br_i the bankruptcy rate.

Year	Purpose	N_i	d_i	p_i	n_i	An_i	br_i
			Belgian data set				
<2013	Train set 1	401,377	41,367	727,008	1,050,206	2.62	0.103
2013	Train set 2	383,841	4,858	721,767	1,058,121	2.76	0.0127
2014	Test set	400,203	6,107	757,407	1,121,321	2.80	0.0153
			UK data set				
<2013	Train set 1	1,824,877	205,617	4,634,296	7,611,938	4.17	0.113
2013	Train set 2	1,836,095	9,746	4,548,715	7,666,828	4.18	0.005
2014	Test set	2,080,127	7,733	4,882,163	8,247,782	3.97	0.004

the observation date) and a label indicating whether they went bankrupt in the course of 2014. Both the Belgian and UK data sets have an increasing number of SMEs between 2013 and 2014, though a decreasing default rate. The Belgian SMEs have a 2 to 3 times higher default rate than the UK companies. Figure 5 gives an example of an actual test node and its links to the companies in the training set. The test node is grey, the companies that defaulted before 2013 are black and the companies that were still active at the end of 2012 are white. The relational model simply links each company of the test set to its neighbours in training set 1 and applies the relational learner to estimate the bankruptcy probability score. Companies that were active in 2012 and 2014 are present in both the network training and test set. For these observations, we apply a similar approach to 'leave-one-out cross-validation', i.e. we build a network on the 2012 training set excluding the values of this company in 2012 and use the company in 2014 as test instance to estimate the bankruptcy probability score.

Figures 6 and 7 facilitate the understanding of our out-of-time ensemble set-up. In the training set, we use the labels for all companies that were active on 1st January 2013. These

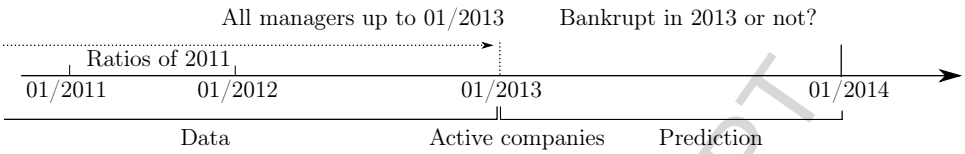


Figure 6: Ensemble training set: we use the labels in 2013 (bankrupt +1 and active 0), the ratios of 2011 and all the managers that are/have been part of the companies that are still active on the prediction date (01/2013).

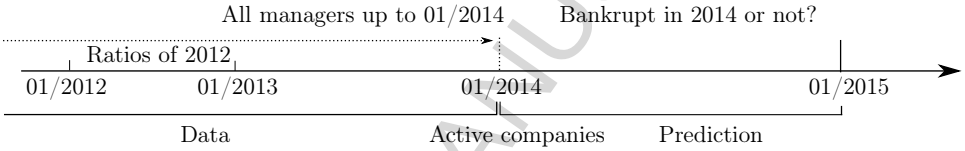


Figure 7: Test set: we want to predict the labels of 2014 (bankrupt +1 and active 0) using the ratios of 2012 and all the managers that are/have been part of the companies that are still active on the prediction date (01/2014).

labels received the value +1 if the company went bankrupt in the course of 2013 and 0 if the company was still active at the end of 2013. Concerning the inputs to our ensemble model, the network scores of the companies are calculated using all managers that were part of the company up to January 2013 and the financial ratios are calculated using the financial statement of 2011[3] Figure 7 illustrates the set-up of our test set. To predict bankruptcy one year ahead for all firms active on 1 January 2014, we use the financial statement of 2012 and the network scores calculated using all managers up to 1 January 2014. A 'leave-one-out' procedure is applied here as well, and is illustrated by Figure 8.[4] We want to predict whether Company 1 in Figure 8 will go bankrupt in the course of 2014. First, we build the

[3]As mentioned earlier, due to the delay on publication, we cannot use the statement of 2012.

[4]To decrease computational time, instead of excluding one company at a time, we exclude companies in chunks of 1000.

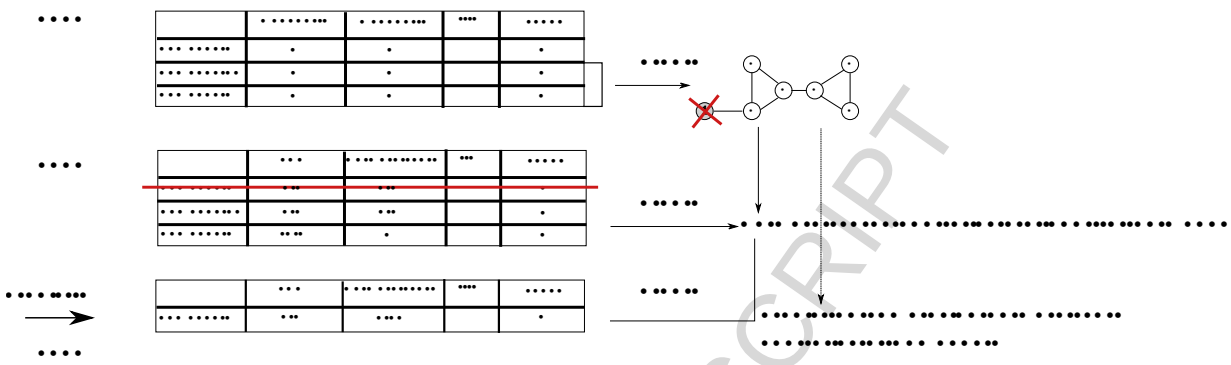


Figure 8: The modelling procedure for the ensemble model when double observations occur.

network using the 'leave-one-out' procedure, i.e. we build a network using the data of 2012, excluding the information on Company 1. This procedure is used for all double occurrences, leading to multiple networks. Next, we train our ensemble model on the 2013 data set. As input features we use the financial ratios and the network scores for all companies except Company 1. The network scores are estimated using the 2012 networks and the updated list of managers in 2013. We finally predict the probability of failure for Company 1 using the ensemble model, the company's financial ratios and network score (estimated using the 2012 network that excludes Company 1 and the updated list of managers for Company 1 in 2014).

The relational bankruptcy probability scores are generated using the smoothed wvRN technique as described in Section 3. For the base and ensemble model, we train an SVM with a linear kernel, a technique that is both powerful and comprehensible, thus rendering it an appropriate choice for the modelling problem at hand. SVM searches for the decision boundary that maximizes the margin between the two classes. The version of Linear SVM used here solves the following optimisation problem [15]:

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i max(1 - y_i\mathbf{w}^T\mathbf{x}_i, 0)^2 \quad (4)$$

where vector $\mathbf{w}$ represents the weights of the model, $\mathbf{x}_i$ and y_i are the input vector and the label of the *ith* observation and $\max(1 - y_i\mathbf{w}^T\mathbf{x}_i, 0)^2$ is the squared (L2) hinge-loss function. An out-of-sample grid search on an in-time, out-of-sample random validation set was performed to find the optimal value of C, the cost parameter. The unbalanced distribution of the classes in the data set, necessitates an undersampling of the negative class (active firms) in the training set for both the base model and the ensemble model. We reduced the non-event rate to 50:50 in the training set since SVM is sensitive to the large class imbalance. The relational learners are a powerful tool and are able to handle the issue of unbalanced classes [19]. The network scores are therefore calculated using the entire network and not a sample. The results are calculated on the complete test set.

5.2. Results

We compare the results of the network, base and ensemble model using the Area under the Receiver Operating Characteristic-curve (AUC) [16] and the lift [11]. In this setting, lift represents the ratio between the bankruptcy rate of a selected group of companies from the test set and the average bankruptcy rate.

Top node weights. Prior to calculating the general results, the appropriate weighting function has to be chosen. We have tested the performance of each top node weighting scheme using ten-fold cross-validation on a balanced sample that contains all bankrupt firms in 2013 and for each fold, a different sample of the non-bankrupt firms in the same year. Every node of the test sample is connected to all their neighbouring nodes in the training set. The results are reported in Table 4. Hyperbolic tangent seems to be the best weighting scheme for both data sets, followed closely by the inverse degree and inverse frequency. These three schemes downweigh top nodes (i.e. directors) that are connected to many firms. Using the time a manager stayed at a company to calculate the edge weight does not add valuable

information to the predictions. In what follows, the hyperbolic tangent function is therefore used to weigh the top nodes.

Table 4: Cross-validated performances of the network model for different top node weighting schemes. The reported results are the AUC and lifts at different percentages, averaged over the ten folds.

Belgian data set					
Weight	AUC	lift 1%	lift 5%	lift 10%	lift 20%
Inverse degree	74.45	1.95	1.93	1.93	1.88
Inverse frequency	74.88	1.75	1.85	1.89	1.86
Hyperbolic tangent	74.59	1.97	1.95	1.93	1.88
Class-degree ratio	73.23	1.85	1.84	1.81	1.80
Delta function	68.68	1.74	1.93	1.92	1.84
Adamic and Adar	68.97	1.76	1.86	1.91	1.84
Time function	72.54	1.77	1.90	1.92	1.88
UK data set					
Weight	AUC	lift 1%	lift 5%	lift 10%	lift 20%
Inverse degree	63.03	1.77	1.70	1.64	1.51
Inverse frequency	62.73	1.70	1.63	1.59	1.47
Hyperbolic tangent	63.03	1.78	1.70	1.64	1.50
Class-degree ratio	61.34	1.69	1.61	1.52	1.46
Delta function	61.41	1.50	1.66	1.60	1.49
Adamic and Adar	63.00	1.51	1.63	1.60	1.48
Time function	61.27	1.54	1.67	1.63	1.49

Main results. The results in Table 5 show that the relational data on its own is insufficient, however with an AUC of 69.56% for Belgium and 66.11% for the UK it still has reasonable predictive power. Adding the network scores to the base model, slightly increases the AUC from respectively 82.86% to 84.71% and 81.29% to 82.68%. However, a much larger increase can be seen at the beginning of the lift curves in Figures 9a and 9b. The ensemble model in Figure 9a has a 37.85 times higher bankruptcy detection rate than the average bankruptcy rate of 1.5% when we consider only the 0.1% highest scores, i.e. the top 400 companies selected. This means that more than half (57.90%) of these companies went bankrupt in

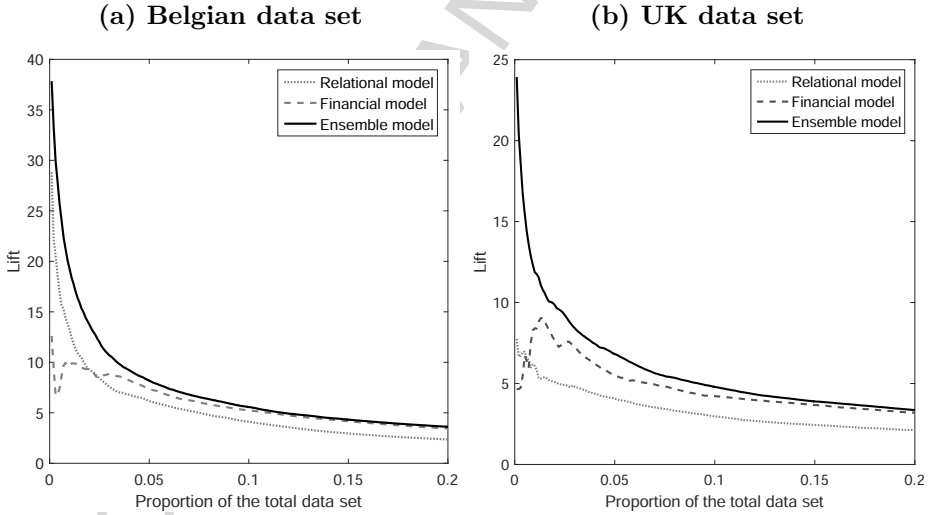
2014. In this segment, the financial model has a detection rate of 19.30%. The ensemble model's lift is comparable with the base model's at the percentiles above 10%. This result confirms that the highest ranked companies in the ensemble model, those connected to many (or only) bankrupt firms, have indeed a higher probability of going bankrupt. However, it also shows that one should still consider their financial circumstances. A similar story goes for the UK data set, as illustrated in Figure 9b. The ensemble model has a 23.93 times higher bankruptcy rate than the average rate when we consider only the 0.1% highest scores, i.e. the top 2000 companies. For this segment, the ensemble model has a 413% higher lift than the base model. Comparable performance between the base model and ensemble model starts at percentiles higher than 15%. The most remarkable difference between the Belgian and UK data set is the performance of the network model. For the 0.1% highest scores, the Belgian network model has a 128% higher lift than the base model, while the UK network model has a lift only 66% higher than the base model. Note that the scaling from these lift curves is different than the lifts reported in Table 4. Lift is dependent on the class distribution. The sample used to select the weights has a 50:50 distribution, while the complete test set has a distribution lower than 99:1. In order to investigate to what extent the way the network information is analyzed, influences the performance, we compare the ensemble models with a financial model that adds, as additional variable, the ratio of the number of previous bankruptcies for all managers/directors to the number of managers/directors. Contrary to our proposed network score, this ratio does not include top node weighting. The lift curves in Figure 10 show that our network score better represents and analyzes the information that is contained in the networks. Interestingly, the linear SVM for the UK assigns a very low weight to the ratio in the finacial+ratio-model, which results in a lift curve that is similar to the financial model's curve. A possible explanation is that the UK allows the use of nominee directors, who act in name of a third person. These nominee directors are often part of a large number of firms, including a large number of default firms. The top node weighting significantly downweighs the influence of these

directors, which apparently leads to better results.

Table 5: Predictive performance on the test set, measured in AUC, of the relational model, financial model and ensemble model for the Belgian and UK data set.

Data set	Relational	Financial model	Ensemble model
Belgium	69.56%	82.86%	84.71%
UK	66.11%	81.29%	82.68%

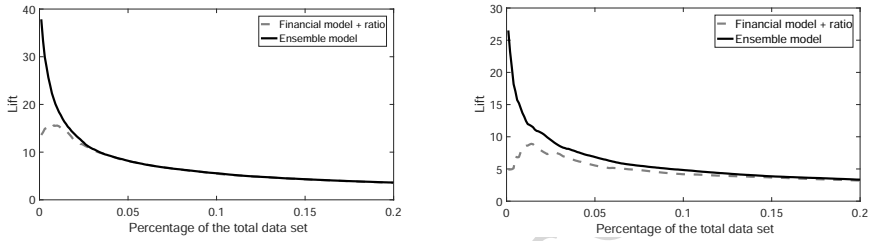
Figure 9: Lift curves of the financial model, relational model and ensemble model for the Belgian and UK data sets.



Ranking of input features. Table 6 shows the ranking of the top five and bottom five (normalised) input features according to the weights of the SVM models. The top five input features have a positive weight and are predictive for bankruptcy. The bottom five features have a negative weight and are predictive for the non-event. For the ensemble model, the most predictive variable for bankruptcy in both data sets is the network score. The base and ensemble model's weights differ slightly in magnitude. This could indicate an interac-

Figure 10: Lift curves for the Belgian and UK data sets of the ensemble models and the financial models that include an unweighted ratio.

(a) Belgian data set **(b) UK data set**



tion between the influence of a company's network and its financial situation. There are, however, no weights that change sign when the network score is added. In both data sets, the Equity ratio is an important predictor, with a missing Equity ratio indicating a higher probability of bankruptcy. The most predictive sectors are different for Belgium and the UK, with only the construction sector appearing in both data set's top rankings.

Additive smoothing. Previous results are created using the smoothed wvRN version. Figure 11 compares the network and ensemble models with and without Laplace smoothing for respectively the UK and Belgian data set. The largest benefit can be found in the Belgian data set, where the network model's lift increases from 3.93 to 28.83 for the first 0.1% companies and from 4.90 to 13.23 for the first 1% companies. A similar trait is present in the UK data.

5.3. Which managers/bankruptcies to include?

In the results so far, we have not taken into account the resignation date of the managers or default date of the companies when creating the bipartite graphs. No distinction is made between managers that are still part of the firm and managers that resigned 10 years ago. Similarly, all bankrupt companies are included in the network. However, it is possible that the influence of a director becomes insignificant some t years after he/she leaves the company and that the influence of a bankrupt firm on its neighbours becomes insignificant some t

Table 6: Ranking of the input features' weights in the ensemble and financial model for the Belgian and UK data set.

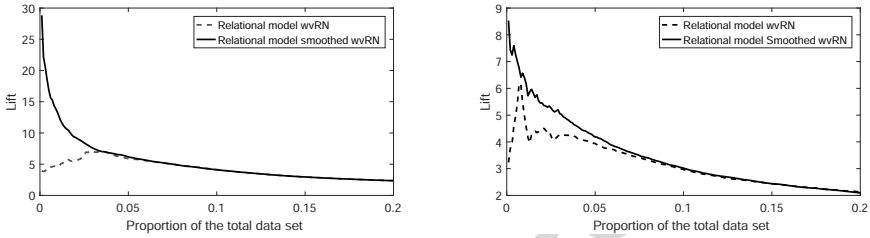
	Belgian data set			
	Financial model	Coefficient	Ensemble model	Coefficient
Positive	Missing ROE	0.8553	Network scores	0.9084
	Missing Equity Ratio	0.3210	Missing ROE	0.8450
	Water supply/sewerage (sector E)	0.1943	Missing Equity ratio	0.3053
	Construction (sector F)	0.1583	Water supply and sewerage (sector E)	0.1519
Negative	Company age	-1.2039	Company age	-1.1210
	Newly founded	-1.0446	Newly founded	-1.0380
	Human health/social work (sector Q)	-0.6036	Human health and social work (sector Q)	-0.6141
	Equity ratio	-0.5963	Equity ratio	-0.5826
	UK data set			
	Financial model	Coefficient	Ensemble model	Coefficient
Positive	Construction (sector F)	0.5437	Network scores	2.2656
	Missing Equity Ratio	0.4962	Missing Equity Ratio	0.4704
	Information and communication (sector J)	0.3868	Construction (sector F)	0.4515
	Missing Current Ratio	0.3468	Cash Flow	0.3492
Negative	Newly founded	-1.4830	Newly founded	-1.4135
	Agriculture (sector A)	-1.1487	Agriculture (sector A)	-1.0957
	Equity Ratio	-1.0508	Equity Ratio	-1.0094
	Activities of extraterritorial bodies (sector U)	-0.7267	Activities of extraterritorial bodies (sector U)	-0.7969

years after the bankruptcy date. To decide which managers and defaults should be included in the bipartite graphs, we have calculated the performance for different values of t on a balanced sample of the 2013 network. For the directors we consider the following options: (i) all the directors of the last t years with $t = \{1, 5, 10, 15, 20\}$ and (ii) all the directors that have ever been part of the firm (before observation date). For the bankruptcies we consider the following options: (i) all the bankruptcies of the last t years with $t = \{3, 5, 10, 15, 20, 25\}$ and (ii) all bankruptcies that are available in the database.

Directors. Table 7 reports the lifts for the different options. Using cross-validation, we obtain the highest 1% lift on the Belgian data set when we include either all the directors or the directors of the last 20 years. However, the difference in both lift and AUC between using all directors or only those of the past 5 years is minimal. Figure 12a shows the results on the complete test set when choosing the directors of the last 5 years, compared to all directors. The lift curves coincide almost completely, with the model using all directors

Figure 11: Lift curves of the smoothed and non-smoothed wvRN network model for the Belgian and UK data sets.

(a) Belgian data set **(b) UK data set**



displaying a higher lift at 0.1%. Regarding the UK data set, the highest lift 1% lift is achieved when we include only the directors of the last 5 years. However, taking into consideration the lift at higher percentages and the AUC-value, the optimal value for t is 10. Including all directors slightly lowers both lift and AUC. Figure 12b shows the results on the test set when selecting the directors of the last 10 years, compared to all directors. The model built using the directors of the last 10 years has a higher lift for the 1% highest scores, however, for higher percentages it is outperformed by the model using all directors.

Bankruptcies. Table 8 reports the lifts for the different options. For both data sets, the highest 1% lift is obtained when the bankruptcies of the last 10 years are included in the network. Slightly higher AUC-values can be found when including older bankruptcies as well, with a maximum AUC-value for $t = 25$, however this value has lower lifts for the reported percentages. Overall, the differences in performance for values of t larger than 5 are minimal. The general conclusion is that, when creating a network of SMEs, the bankruptcies of at least the last 10 years should be included. Figures 13a and 13b compares the performance of the complete network and the network that is restricted to the bankruptcies of the last 10 years. For both data sets, the lift curves of the complete and restricted networks coincide for most points on the graph, thus confirming that it is unnecessary to include bankruptcies older than 10 years.

Table 7: Cross-validated performances of the relational models for different selections of directors using a sample of training set 2. The inclusion of a link between a director and a company in the bipartite graph depends on the time between the observation date and the resignation date.

Belgian data set					
Years since resignation	AUC	lift 1%	lift 5%	lift 10%	lift 20%
$\leq$ 1 year	74.39	1.94	1.91	1.90	1.86
$\leq$ 5 years	74.57	1.96	1.94	1.92	1.88
$\leq$ 10 years	74.58	1.96	1.95	1.93	1.88
$\leq$ 15 years	74.58	1.96	1.95	1.93	1.88
$\leq$ 20 years	74.58	1.97	1.95	1.93	1.88
All directors	74.59	1.97	1.95	1.93	1.88
UK data set					
Years since resignation	AUC	lift 1%	lift 5%	lift 10%	lift 20%
$\leq$ 1 year	61.32	1.72	1.65	1.63	1.40
$\leq$ 5 years	62.56	1.81	1.70	1.64	1.49
$\leq$ 10 years	63.17	1.79	1.70	1.65	1.51
$\leq$ 15 years	63.14	1.79	1.67	1.64	1.51
$\leq$ 20 years	63.05	1.79	1.70	1.64	1.50
All directors	63.03	1.78	1.70	1.64	1.50

Figure 12: Lift curves of the complete and restricted networks for the Belgium and UK test sets. The restricted networks contain only the directors of the last 5 years for the Belgian data set and the last 10 years for the UK data set.

(a) Belgian data set **(b) UK data set**

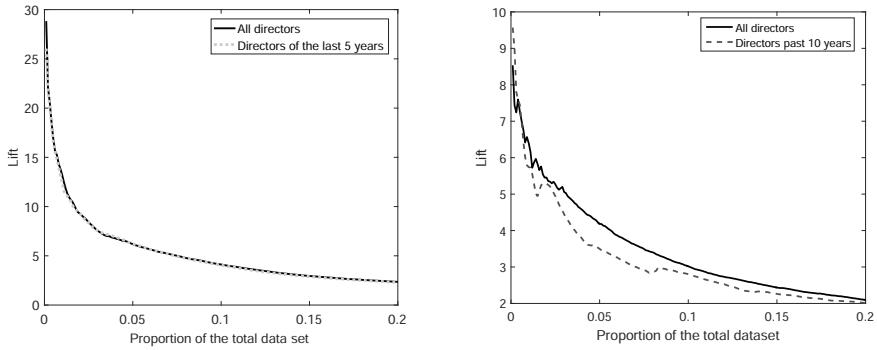
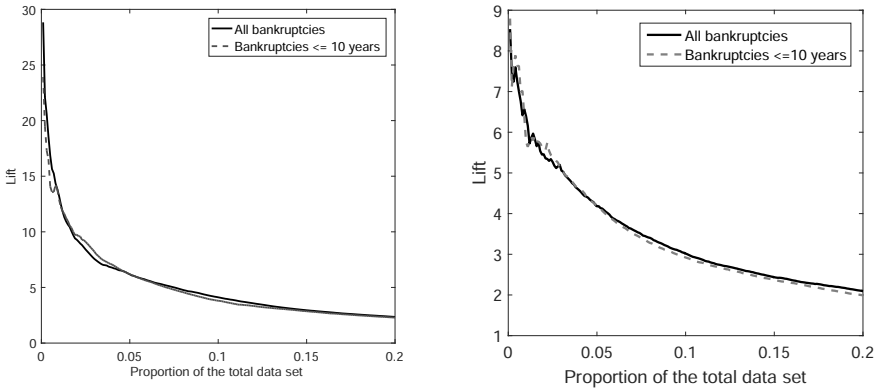


Table 8: Cross-validated performances of the relational models for different selections of bankrupt firms using a sample of training set 2. The inclusion of a bankrupt firm in the bipartite graph depends on the time between the observation date and the bankruptcy date.

Belgian data set					
Years since bankruptcy	AUC	lift 1%	lift 5%	lift 10%	lift 20%
$\leq$ 3 years	65.47	1.94	1.95	1.95	1.77
$\leq$ 5 years	72.02	1.96	1.95	1.95	1.90
$\leq$ 10 years	74.57	1.97	1.96	1.94	1.90
$\leq$ 15 years	74.60	1.97	1.95	1.94	1.88
$\leq$ 20 years	74.63	1.97	1.95	1.93	1.88
$\leq$ 25 years	74.62	1.97	1.95	1.93	1.88
All bankruptcies	74.59	1.97	1.95	1.93	1.88
UK data set					
Years since bankruptcy	AUC	lift 1%	lift 5%	lift 10%	lift 20%
$\leq$ 3 years	57.15	1.71	1.62	1.47	1.30
$\leq$ 5 years	60.68	1.81	1.72	1.64	1.43
$\leq$ 10 years	62.85	1.82	1.72	1.67	1.51
$\leq$ 15 years	62.91	1.81	1.72	1.67	1.51
$\leq$ 20 years	62.94	1.80	1.72	1.65	1.51
$\leq$ 25 years	63.03	1.78	1.70	1.64	1.51
All bankruptcies	63.03	1.78	1.70	1.64	1.50

Figure 13: Lift curves of the complete and restricted networks for the Belgian and UK test sets. The restricted networks include the bankruptcies of the last 10 years only.

(a) Belgian data set **(b) UK data set**



5.4. Deployment and limitations

A major advantage of the proposed design is that it utilizes a data source that is often already at a bank's disposal. The model can serve multiple purposes: it can be used as by financial institutions as credit scoring model or to estimate the credit risk parameters, it can help banks manage their credit line as a suddenly high network score might be interpreted as a warning flag of possible pending bankruptcy, and it can assist or replace the decision logic phase. During the credit scoring process, credit scores are validated by the bank's decision logic. During this phase, applicants with a positive credit score can still be retained by the bank if they are deemed risky for reasons supported by expert knowledge. One of the reasons could be a link to a fraudulent or bankrupt firm, or family ties with a persistent defaulter. It is thus possible that the companies that are detected by our ensemble and network model, would be detected by the bank during the decision logic phase. In this case, the added value of our methodology is the automation of the decision logic process. Since we did not have the necessary information at our disposal, we could not test whether we found companies that would not be detected by the bank at any point during the scoring process. However, the 400 highest scoring companies of the ensemble model have an average network score of 0.477, possibly indicating that these are not the firms with the obvious networks of being connected to only bankrupt firms.

One of the major limitations of the proposed models is that, due to the ensemble set-up, we lose information. Moreover, the model does not account for simultaneous bankruptcies, nor bankruptcies that occur in the same year. Another issue is that the model building is rather slow because of the leave-one/many-out procedures [5]. However, in reality banks will have a lower amount of firms that they need to make predictions for. Finally, a major limitation is that the models might come with increased model risk. Model risk can be

[5] Using Matlab on an Intel Core i5-3470 CPU @ 3.20 GHz machine with 8Gb RAM, it took approximately 26 hours to create network scores for the UK test set and 2 hours for the Belgian test set. Building the ensemble model took 1 hour for the Belgian test set and 6 hours for the UK test set.

defined as the risk that a model's outcomes are systematically wrong [36]. As the models come with increase model complexity, the results of stress tests or backtests (comparing the model's outcomes to realized values) might be more difficult to interpret. Moreover, the model that we propose is highly dependent on the population and therefore its performance might deteriorate when the population drifts from the reference data set[36]. Employing a novel model in credit scoring should thus be accompanied by rigorous sensitivity analyses.

6. Conclusion

In this paper, we report the potential of relational data for bankruptcy prediction using two large, real-life SME data sets. We show that linking companies based on their managers/board members adds complementary predictive power to the traditional bankruptcy prediction. The results confirm the large predictive value of relational data and demonstrate that this mostly unused data source should be considered when developing bankruptcy prediction models. The proposed design can be easily implemented by financial institutions and credit rating bureaus as this data source is often already at their disposal. Moreover, the smoothed wvRN does not require large IT infrastructures. The methodology can be extended to different applications in banking, such as loan default prediction, fraud detection and marketing. Additionally, the design can be helpful in B2B commerce for targeted advertising and churn prediction.

7. Acknowledgements

The authors thank the Flemish Research Foundation (FWO) for their financial support (Grant number G.0827.12N and the personal grant for <name removed>).

References

[1] Ahn, B., Cho, S., and Kim, C. (2000). The integrated methodology of rough set theory and artificial neural network for business failure prediction. *Expert Systems with Applications*, 18(2):65–74.

[2] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609.

[3] Altman, E. I. and Sabato, G. (2007). Modelling credit risk for smes: Evidence from the us market. *Abacus*, 43(3):332–357.

[4] Altman, E. I., Sabato, G., and Wilson, N. (2010). The value of non-financial information in small and medium-sized enterprise risk management. *The Journal of Credit Risk*, 6(2):95.

[5] AYADI, R. (2005). The New Basel Capital Accord and SME Financing. *Center for European Policy Studies, Brussels.*

[6] Baldwin, J. R. (1998). Failing concerns: business bankruptcy in Canada. *Failing Concerns: Business Bankruptcy in Canada.*

[7] Basel Committee on Banking Supervision (2006). International convergence of capital measurement and capital standards: a revised framework. Technical report, Bank for International Settlements.

[8] Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, pages 71–111.

[9] Bell, T. B. (1997). Neural nets or the logit model? a comparison of each model's ability to predict commercial bank failures. *Intelligent Systems in Accounting, Finance and Management*, 6(3):249–264.

[10] Bellovary, J. L., Giacomino, D. E., and Akers, M. D. (2007). A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial Education*, pages 1–42.

[11] Berry, M. J. and Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management.* John Wiley & Sons.

[12] Bian, H. and Mazlack, L. (2003). Fuzzy-rough nearest-neighbor classification approach. In *Fuzzy Information Processing Society, 2003. NAFIPS 2003. 22nd International Conference of the North American*, pages 500–505. IEEE.

[13] Bocken, H. and De Bondt, W. (2001). *Introduction to Belgian law*. Kluwer Law International.

[14] Davis, A. and Veloso, A. (2016). Subject-related message filtering in social media through context-enriched language models. In *Transactions on Computational Collective Intelligence XXI*, pages 97–138. Springer.

[15] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

[16] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

[17] Goode, R. (2011). *Principles of corporate insolvency law*. Sweet & Maxwell.

[18] Hill, S., Provost, F., and Volinsky, C. (2006). Network-based marketing: Identifying likely adopters via consumer networks. *Statist. Sci.*, 21(2):256–276.

[19] Junqué de Fortuny, E., Stankova, M., Moeyersoms, J., Minnaert, B., Provost, F., and Martens, D. (2014). Corporate residence fraud detection. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1650–1659. ACM.

[20] Lee, K. C., Han, I., and Kwon, Y. (1996). Hybrid neural network models for bankruptcy predictions. *Decision Support Systems*, 18(1):63–72.

[21] Ma, Y., Ansell, J., and Andreeva, G. (2014). Exploring management capability in SMEs using transactional data. *Journal of the Operational Research Society*, 67(1):1–8.

[22] Macskassy, S. A. and Provost, F. (2007). Classification in networked data: A toolkit and a univariate case study. *The Journal of Machine Learning Research*, 8:935–983.

[23] Martens, D. and Provost, F. (2011). Pseudo-social network targeting from consumer transaction data. Technical Report CEDER-11-05, New York University.

[24] McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, pages 415–444.

[25] Min, J. H. and Lee, Y.-C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, 28(4):603–614.

[26] National Bank of Belgium (2015). Central credit register: total of credits granted to resident non-financial corporations. [Online; accessed 3-July-2015].

[27] Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, pages 109–131.

[28] Olson, D. L., Delen, D., and Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, 52(2):464–473.

[29] Ooghe, H. and De Prijcker, S. (2008). Failure processes and causes of company bankruptcy: a typology. *Management Decision*, 46(2):223–242.

[30] Ooghe, H. and Van Wymeersch, C. (2008). *Handboek financiële analyse van de onderneming*. Intersentia nv.

[31] Sharma, S. and Mahajan, V. (1980). Early warning indicators of business failure. *The Journal of Marketing*, pages 80–89.

[32] Shin, K.-S., Lee, T. S., and Kim, H.-j. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, 28(1):127–135.

[33] Stankova, M., Martens, D., and Provost, F. (2015). Classification over bipartite graphs through projection. Technical report, University of Antwerp Working Paper.

[34] Tinoco, M. H. and Wilson, N. (2013). Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis*, 30:394–419.

[35] Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67.

[36] Van Gestel, T. and Baesens, B. (2009). *Credit Risk Management: Basic concepts: Financial risk components, Rating analysis, models, economic and regulatory capital*. Oxford University Press.

[37] Van Gestel, T., Baesens, B., Suykens, J., Espinoza, M., Baestaens, D.-E., Vanthienen, J., and De Moor, B. (2003). Bankruptcy prediction with least squares support vector machine classifiers. In *Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on*, pages 1–8. IEEE.

[38] Weber, I., Garimella, V. R. K., and Borra, E. (2013). Inferring audience partisanship for youtube videos. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 43–44. International World Wide Web Conferences Steering Committee.

[39] Wilson, R. L. and Sharda, R. (1994). Bankruptcy prediction using neural networks. *Decision Support Systems*, 11(5):545–557.

[40] Zhang, G., Hu, M. Y., Patuwo, B. E., and Indro, D. C. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European Journal of Operational Research*, 116(1):16–32.

Biography

Ellen Tobback is a PhD student at the University of Antwerp, Belgium, where she received her Master's and Bachelor's degrees in Commercial Engineering. Since 2013 she has been a member of the Applied Data Mining research group at the Department of Engineering Management. She received a PhD Fellowship from the Flanders Research Foundation and expects to complete her PhD dissertation in 2017. Her general research area is data mining for credit risk prediction.

Tony Bellotti is a senior lecturer in statistics in the Mathematics department at Imperial College London. He received his PhD in computational learning from Royal Holloway, University of London in 2006 and his main research interest is the application of statistical and machine learning models in credit risk analysis. He is an Honorary Fellow in the Credit Research Centre at the University of Edinburgh.

Julie Moeyersoms graduated in 2012 as business engineer at the Faculty of Applied Economics at the University of Antwerp. In April 2016, she finished her PhD at the Applied Data Mining group in the Department of Engineering Management. The topic of her dissertation is on the (predictive and economic) value of customer behavior data in predictive modeling. Several real life data sets are researched, going from marketing applications such as churn prediction or online advertising, to risk management applications such as residence fraud detection and credit risk predictions.

Marija Stankova graduated in 2012 as computer engineer at the Ss. Cyril and Methodius University of Skopje. In December 2016, she finished her PhD at the Applied Data Mining group at the University of Antwerp. The topic of her dissertation is on classification within network data with a bipartite structure. Her research has successfully been applied to real-life data sets, such as corporate residence fraud and microfinance lending.

David Martens is assistant professor at the University of Antwerp, where he heads the Applied Data Mining research group. His research focuses on the development and application of data mining techniques that lead to an improved understanding of human behavior, and their use in marketing and finance. In 2014, David won the ''Best EJOR Application Award'' (European Journal of Operational Research) for his work on churn prediction. In 2008, David was finalist of the prestigious international KDD doctoral dissertation award. Together with prof. Foster Provost, he is the inventor of four pending patent applications of data mining methods.

Highlights

- A new SME bankruptcy prediction model that includes relational data is proposed.
- The model links two companies using shared directors and managers.
- A relational classifier is applied to the resulting network.
- Relational data helps detecting the riskiest firms.
- Relational and financial data have complementary predictive power.