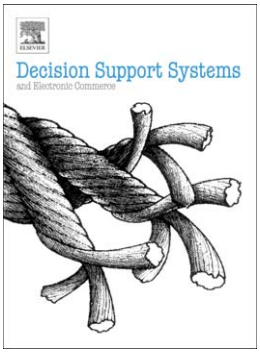


Predicting tax avoidance by means of social network analytics

Jasmien Lismont, Eddy Cardinaels, Liesbeth Bruynseels, Sander De Groote, Bart Baesens, Wilfried Lemahieu, Jan Vanthienen

PII: S0167-9236(18)30022-8
DOI: doi:10.1016/j.dss.2018.02.001
Reference: DECSUP 12927

To appear in: *Decision Support Systems*

Received date: 19 July 2017
Revised date: 30 January 2018
Accepted date: 1 February 2018

Please cite this article as: Jasmien Lismont, Eddy Cardinaels, Liesbeth Bruynseels, Sander De Groote, Bart Baesens, Wilfried Lemahieu, Jan Vanthienen, Predicting tax avoidance by means of social network analytics, *Decision Support Systems* (2018), doi:10.1016/j.dss.2018.02.001

Predicting tax avoidance by means of social network analytics

Jasmien Lismont[a,*], Eddy Cardinaels[b,c], Liesbeth Bruynseels[b], Sander De Groote[b], Bart Baesens[a,d], Wilfried Lemahieu[a], Jan Vanthienen[a]

[a] *KU Leuven, Dept. of Decision Sciences and Information Management, Naamsestraat 69, B-3000 Leuven, Belgium*
[b] *KU Leuven, Dept. of Accountancy, Finance & Insurance, Naamsestraat 69, B-3000 Leuven, Belgium*
[c] *Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands*
[d] *University of Southampton, University Road, Southampton SO17 1BJ, United Kingdom*

Abstract

This study predicts tax avoidance by means of social network analytics. We extend previous literature by being the first to build a predictive model including a larger variation of network features. We construct a network of firms connected through shared board membership. Then, we apply three analytical techniques, logistic regression, decision trees, and random forests; to create five models using either firm characteristics, network characteristics or different combinations of both. A random forest including firm characteristics, network characteristics of firms and network characteristics of board members provides the best performance with a minimal increase of 7 pp in AUC. Hence, including network effects significantly improves the predictive ability of tax avoidance models, implying that board members exhibit specific knowledge which can carry over across firms. We find that having board members with no connections to low-tax companies lowers the

*Corresponding author
Email address: Jasmien.Lismont@kuleuven.be (Jasmien Lismont)

likelihood of being a low-tax firm. Similarly, the higher the average tax rate of the companies a board member is connected to, the lower the chance of being low-tax. On the other hand, being connected to more low-tax firms increases the probability of being low-tax. Consistent with prior literature on firm-specific variables, PP&E has a positive influence on the probability of being low-tax, while EBITDA has a negative effect. Our results are informative for companies as to the director expertise they want to attract in their boards. Additionally, financial analysts and regulatory agencies can use our insights to predict which firms are likely to be low-tax and potentially at risk.

Keywords: board interlocks, predictive analytics, social network analytics, social ties, tax avoidance, tax planning

---

1. Introduction

There is considerable variation in taxes being paid among corporate organizations [20]. While firms enjoy benefits of tax avoidance by lower taxes being paid, it does not come without risk as authorities may impose fines and penalties for tax evasion, and tax avoidance may involve significant political and reputational costs [27]. Motivated by the variation in taxes being paid and the different trade-offs for tax avoidance, researchers start to examine the determinants of why companies engage in tax avoidance. In this context, many studies focus on firm-specific variables and the various incentives that managers receive [2, 14, 33, 38]. In addition, one may look at the different governance variables, the quality of information systems, and various types of expertise in the board or audit office, indicating that tax planning does require a certain level of expertise [17, 27, 32]. This paper uses techniques from the social network analytics domain to develop a predictive

classification model for tax avoidance. Motivated by recent findings that companies acquire the expertise required for the successful execution of corporate strategies through their network of directors [5, 19, 23, 25], we zoom in on several firm and director network features and their predictive validity in explaining tax avoidance. Up to present date, not much is known on the role of social linkages of directors across firms and whether they share crucial knowledge that can explain tax avoidance (20, p. 146). We look at how firms are connected through shared directorships and how shared knowledge in the network via connections to low-tax firms and non-low-tax firms can be informative for making predictions on whether a firm will be able to maintain (or become) a low-tax firm in the future. We show that a combination of firm characteristics and network characteristics of both the firm and its board members provides the best predictive performance. As such, a hybrid model including both firm and network characteristics (using a random forest) is able to identify more low-tax firms, and highlights the importance of several network features in predicting tax avoidance.

Our study delivers several research and managerial contributions. Firstly, by including an extensive set of network features and by building a more complex network, we are able to benefit from more detailed information about board members and their relations to other companies. This improves the comprehensibility of our models, since the features allow for a discussion of the impact of specific network characteristics. Furthermore, the predictive performance of our models is improved which allows for a better identification of low-tax firms. Secondly, we contribute to management and society. Our results inform the management about the director expertise they want to attract if they desire a low-tax strategy. They confirm that attracting directors who are connected to low-tax firms now or in

the past, can affect the companies' own tax rate, suggesting that these directors deliver crucial knowledge or have valuable expertise for maintaining a low-tax strategy. Additionally, our predictive models can aid in a priori identification of firms that will maintain a low-tax strategy in the future and as such provide valuable insights for financial analysts and regulatory agencies.

Our paper is structured as follows. Firstly, in Section 2, we discuss related research on tax avoidance and social network analytics to illustrate the importance and novelty of our study. Next, Section 3 describes our methodology. Our results are presented in Section 4 and consecutively discussed. Finally, Section 5 concludes our study.

2. Related research

Previous studies illustrate that human actors in firms have access to specific resources and that the knowledge of these directors, auditors or law firms travels across their network. Han et al. [19] study the effect of director interlocks on corporate R&D investments. They found that managers imitate the R&D investment intensity of their interlocked firms. Horton et al. [23], Larcker et al. [28]; and Omer et al. [35] take a closer look at firm performance and how directors' connectedness impacts this. With regards to company revenue relations prediction, Ma et al. [30] develop a network of firms based on citations in news stories. They focus on centrality measures as well as the PageRank and the HITS algorithm, known for web page ranking purposes. Centrality measures provide information on how the entity is positioned in the network. Schabus [40] concludes that the management forecast of earnings from firms with better connected directors, are much

more accurate. In earnings management, social networks may also have an effect. Chiu et al. [11] indicate that earnings management contagion occurs more often for firms who have directors in common. Furthermore, Bizjak et al. [5] show that firms who have a board member coming from a backdating firm, are more likely to backdate stock options themselves. In the same context, Dechow & Tan [12] discovered that backdating firms are more highly connected via shared law firms. These studies [23, 28, 35, 40] include in general a combination of centrality network features, such as degree, closeness, betweenness and eigenvector centrality. Others [5, 11, 12] include network features by calculating the number of links to other firms experiencing the outcome variable.

Following this line of reasoning, researchers start to look at the impact of network effects on tax avoidance. A network then consists of either companies or directors that are linked or connected to each other. For example, companies can be linked because they share common resources, such as board members, auditors, law firms, executives, etc. Directors and executives alike can be connected because they sit on the same board, share their job title, or know each other in a social context [8, 19, 36]. Dyreng et al. [16] examine, for example, whether executive effects, next to firm characteristics, impact tax avoidance. Tracking individual executives across companies, they show that executives play a pivotal role in the level of a company's tax avoidance behavior. The authors only look at characteristics of the individuals and do not take network effects of these executives into account. Nevertheless, their results hint at the fact that such network effects can be important next to firm characteristics. Bianchi et al. [4] found that better connected auditors have an impact on their clients' tax avoidance by including degree, eigenvector and betweenness centrality measures. Neuman

[34] includes directors' connections in order to gain insight into firms' tax planning. For this purpose, they extract four centrality features from a social network of directors, namely degree, betweenness, closeness and eigenvector centrality. Brown & Drake [7] examine the impact of board interlocks on tax avoidance rates by extracting the number of ties to low-tax firms. They illustrate that firms who have more board members tied to low-tax firms, enjoy lower tax rates themselves. Jiang et al. [25] focus on how well the focal firm is connected to firms in well-known islands considered as tax havens, to describe current tax avoidance behavior of companies.

We contribute to this domain by being the first in an accounting context to develop a more extensive set of network measures which are validated by means of advanced machine learning techniques [29]. Previous literature, in general, incorporates network effects parsimoniously by means of two types of variables. Either they apply non-network characteristics, e.g. by focusing on the directors, or they focus solely on centrality measures which indicate the position of an entity, e.g. a director, in the network and how close this entity is to others as measured by degree, closeness, or betweenness centrality. Moreover, these features sets are frequently limited in size and often do not contain more than three network variables. We offer a broader picture of which network features are informative for tax planning activities of firms. Furthermore, we study the effect of including bipartite network features. This means that we are able to include characteristics of firms as well as directors and the explicit connections between them. This allows for the inclusion of more information, in particular because we allow that these connections are weighted by the strength of the relation of the board member to the company. Next, the notable exceptions investigating the effect of social networks for tax avoidance [4, 7, 34, 25] do this in a more

descriptive manner. They focus on how well one firm is connected to other firms via shared directorship to describe current tax avoidance behavior of companies without modeling any predictive features. We extend this research by, to the best of our knowledge, being the first to build a predictive model for tax avoidance and thus providing insights on the predictive value of several network features, relative to firm-specific variables. This offers new perspectives on the economic importance of network effects in the tax planning of firms, and on the validation of the predictive value of different social network techniques in the domain of tax planning [3; 22, p. 246].

Our predictive models are of interest to management, shareholders, and directors that are involved in the tax planning strategies for the company [18]. Shareholders can benefit from aggressive tax policies and a low tax rate. Companies rank increased earnings per share as one of the key reasons for engaging in tax planning activities [18]. Additionally, management and corporate directors (including tax directors) often receive significant financial incentives which may further increase the motivation to engage in aggressive tax planning [2, 38]. Such parties may be interested in the impact of network variables on taxes being paid. Attracting knowledgeable board members from other low-tax firms, may be beneficial to the own corporate company and executives may use their influence to appoint these types of directors. Secondly, our models can inform intermediaries (e.g. financial analysts) who either assess the firm's risk, or tax authorities that want to target firms for investigation. Aggressive tax avoidance, as proxied by firms who can consistently sustain low effective tax rates compared to their peers, raises risks for investors of the companies [38]. Such companies may face higher public scrutiny [21]. Financial analysts can incorporate this risk better, based on the parameters we predict to be crucial for tax avoid-

ance. Additionally, as noted by Slemrod [41], US regulators increased their focus on the related topic of tax evasion after the financial crisis of 2008 both in terms of policy and enforcement. The Internal Revenue Service also uses modern data analytics techniques to identify potential firms to target. Our results provide unique insights to identify the crucial variables that are likely to predict whether a company would be a low-tax firm in the future and thus help the tax authorities to better target their resources towards firms that are more likely to engage in aggressive tax policies. They, furthermore, add value to these agencies as we highlight the importance of —potentially overlooked— network effects. We accomplish this by, for example, suggesting that when managers of firms are linked to firms that have higher average tax rates, or when connections to low-tax firms are absent, such firms deserve less investigation compared to firms that are connected to low-tax firms.

3. Methodology and feature extraction

By means of predictive analytics models, we aim to classify a firm as low-tax or not. Section 3.1 provides more information about the general set-up and data sources used, see Figure 1. Then, we discuss how the dependent variable is measured, and cover how the network is built. Next, we discuss how we extracted the features employed in the models in Section 3.4 and how we compare the different models.

3.1. Methodology

We build predictive models by including both firm characteristics, to which we will refer as local features, and network characteristics. These network characteristics indicate how knowledge can be transferred between

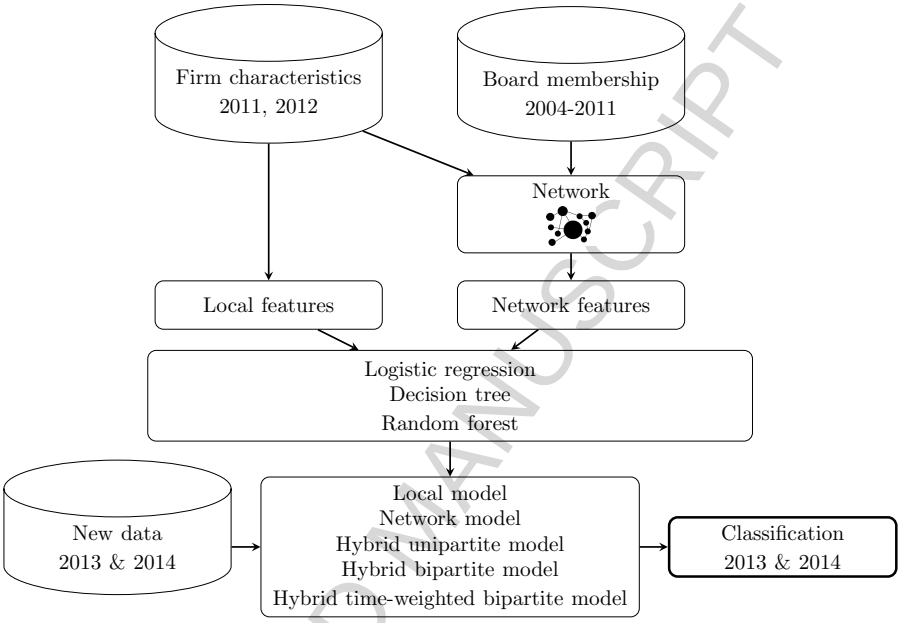


Figure 1: Methodology

firms by means of shared board members. For this purpose we start from two datasets. On the one hand, we have collected firm characteristics data of 1,032 firms from Compustat for fiscal year 2011 and their tax avoidance data for 2012 because we want to identify which companies are going to be low-tax next year by using data of the current year. The lion's share of the companies are listed firms located in the US, and operating in various sectors such as durable goods, retail, computers, services, financial, transport, textiles, etc. Table 1 summarizes the firms characteristics for the employed training set. On the other hand, we extracted data on 42,298 corporate board members from BoardEx for fiscal years 2004 until 2011. This information gives use the capability of creating a network of companies and their board members. Then, we extract both local features and network features from the respective datasets to form our training set for the predictive models.

Table 1: Summary of firm characteristics for the firms in the training dataset.

Feature	**[Minimum, Maximum]**	**Average**
EBITDA	[0.002868,0.6790]	0.1818
R&D	[0,0.4566]	0.02317
Advertising	[0,0.2279]	0.01176
SG&A	[0,0.8419]	0.1934
Capex	[0,0.7753]	0.1095
Sales	[-0.7815,6.5757]	0.1583
Leverage	[0,3.0201]	0.2174
Cash	[0,0.8560]	0.1464
FOR	no: 441 (43%) & yes: 591 (57%)	
NOL	no: 509 (49%) & yes: 523 (51%)	
Size	[2.398,11.647]	7.748
Intangibles	[0,0.7612]	0.2302
PP&E	[0.002744,2.1744]	0.4605

In the training dataset, 9.11% are low-tax firms while 17.83% are high-tax firms. By means of three analytics techniques, namely logistic regression, decision trees, and random forests, we build five models to compare. Logistic regression and decision trees are common techniques for classification tasks. Random forests are an ensemble technique which constructs multiple decision trees and combines them into one model, and is believed to deliver superior performance [29]. We focus on multiple techniques in order to provide different points of view on predictive models.

The five models are distinctive based on the feature sets they use as input. The first model is a local model using only firm characteristics. This model can be regarded as the current state and as a benchmark against which we can compare the other models. The network model uses only network characteristics and the hybrid models use different combinations of local and network characteristics as we will discuss in Section 3.4. We consecutively use these five models to make predictions on new data from the following two

Figure 2: Data collection for training and testing.

years. Performance is thus compared on two out-of-time test sets, namely for 1,251 firms from fiscal years 2013 and 2014. For this purpose, the firm characteristics are taken from 2012 and the board membership data from 2004 until 2012 but the tax avoidance rates are taken from 2013 and 2014, similarly to the training set. This process is depicted in Figure 2. Also for the test sets we can calculate the low- and high-tax ratios. In 2013, 11.03% of the firms have a low tax rate while 17.91% have a high tax rate. Similarly, in 2014, 10.31% are low-tax firms and 20.78% are high-tax firms. Furthermore, we take a look at how the tax rates of the original 1,032 firms change over time. As such, we discover that from 2012 to 2013 8.04% of the firms changed their tax rate level (low versus not-low) and from 2013 to 2014 7.07% changed.

3.2. Dependent variable: tax avoidance

The tax rate of each firm is based on a three-year average measure of cash effective tax rates (CETR) as defined by Brown & Drake [7], see Equation 1, with i referring to firm i; p indicating the rolling three-year period within the time frame; TXPD are the cash taxes paid; PI is the pre-tax income; and SPI are the special items. We focus on cash ETR because Neuman [34] claims this measure is more representative and comprehensive for a firm's

tax planning strategy. Moreover, we employ a three year window to filter out the natural variance in cash tax rates and as such identify the firms who successfully follow a low-tax strategy [15]. Measuring tax avoidance over a longer time frame allows us to see through the natural variation in corporate tax rates and single out firms who are effectively following a low-tax strategy [7, 15].

$$\text{CETR}_{i,p} = \sum_{t=1}^{3}(\text{TXPD}_{i,t}) / \sum_{t=1}^{3}(PI_{i,t} - SPI_{i,t}) \tag{1}$$

Next, we identify low-tax firms as firms ranked in the lowest quintile based on CETR and adjusted for industry mean as suggested by Brown & Drake [7]. Similarly, high-tax firms are distinguished. We specifically focus on categorization instead of predicting a continuous tax rate to single out firms successfully following a low-tax strategy compared to their industry peers. This helps us to identify those firms with access to the knowledge and resources needed to follow such a strategy. In addition, corporate governance effects are stronger for more extreme forms of tax avoidance [1] allowing us to identify which variables are affecting the strong end of the tax avoidance scale. Moreover, since our goal is to identify companies maintaining a low tax rate, external analysts and regulatory agencies can use our model to focus only on the most tax-aggressive companies. As such, they are able to focus their resources on those companies which are most likely to be at risk.

3.3. Building a social network

A network, consisting of nodes and the edges that link them, is represented by means of a graph. We can distinguish graphs based on the different types of nodes they have, i.e. unipartite (with only one type of

node) or multipartite graphs. In this problem setting, we work with both a unipartite and a bipartite graph. In the unipartite graph, firm nodes are connected with each other if they have current or previous board members in common. In the bipartite graph, we have firm and board member nodes. Each board member is connected with one or more firms and the other way around. Bipartite graphs have the advantage to be more detailed since we can also include director-firm information on top of firm-firm information. In addition, we can assign weights to edges to scale the strength of the connection.

We create three types of graphs which differ based on their network structure and wwith different levels of complexity. (1) Firstly, we create a unipartite, undirected, weighted graph, where each edge is assigned a weight based on the number of shared board members. (2) Secondly, we create a bipartite, undirected, unweighted graph. (3) Thirdly, we create a bipartite, undirected, time-weighted graph. We start from the same setup as in the second graph but we weigh each edge by the membership of this specific board member in time. As such, board members who are currently sitting on a board receive a weight of 1 for this edge. If they have already left this firm, the weight of their connection diminishes just like we assume it does in reality. The weight W is then represented by Equation 2 based on Van Vlasselaer et al. [43]. The decay factor γ is set to 0.6, and was determined based on the time frame of our training dataset running from fiscal years 2004 until 2012. h represents the number of years the board member is not sitting on the board any longer with $h = 0$ for current board members.

$$W_{i,j} = e^{-\gamma h} \quad \text{if a relationship exists between firm } i \text{ and board member } j$$
$$W_{i,j} = 0 \quad \text{otherwise} \tag{2}$$

3.4. Feature selection and extraction

Local variables which represent the firm characteristics are based on the definitions of Dyreng et al. [16] and can be found at the top of Table 2. All local variables are winsorized at the 1% level to reduce the effect of outliers.

Next, there are multiple ways to use network characteristics in an analytical model [31]. We chose to extract features from the network so that we are able to use them by non-relational predictive analytics techniques, such as logistic regression and decision trees. Moreover, this allows us to analyze the effects of the network features. This process is also referred to as featurization or propositionalization [26]. Table 2 presents network features related to tax avoidance (i.e. connections to low- and high-tax firms), which we deduct from the firm's network along with their descriptions. In this table, we refer to first and second order neighbors. The former defines the immediate neighbors a firm is connected to in the network. In the unipartite network, these are the companies with whom the firm of interest shares board members with (currently or in the past). Second order neighbors refer to neighbors who are two steps away from the firm of interest. This is particularly interesting for the bipartite graphs because here firms are only connected to board members. In this case, a second order neighbor is a firm which is connected to a board member of the firm of interest. Furthermore, we use the concept of triangles as suggested by Van Vlasselaer et al. [44] in a fraud detection context. A triangle (see features LowTri, NLowTri and

RLowTri in Table 2) is a closed triplet in the neighborhood of the firm of interest. However, in the bipartite networks it is not possible to discover triangles since no two firms are directly connected to each other. Therefore, we take a look at some characteristics in the network of the board members themselves, see features LowBM, NLowBM, CETRBM and Busy and their weighted counterparts in Table 2. Note that the betweenness was not calculated for the nodes in the bipartite graphs due to the large computation efforts for this measure. Furthermore, weighted features, e.g. WLowdegree, can only be calculated for weighted graphs.

Finally, we combine these local and network features in five feature selection variants. The specific variables included in each model are depicted in Table 2. As such, the first model is a local model which only uses local characteristics, also referred to as firm characteristics. The second model only uses network characteristics from the unipartite network and is referred to as the network model. Thirdly, we construct a hybrid model using both local variables and network variables extracted from the unipartite network. Similarly, models four and five combine local variables with network variables from the unweighted and time-weighted bipartite network respectively.

Table 2: Local and network variables and their description. Columns L; N; HU; HB; and HBT indicate whether the variable is considered for respectively the local; unipartite network; hybrid unipartite network; hybrid unweighted bipartite network; and hybrid time-weighted bipartite network model.

Variable	**Description**	**L**	**N**	**HU**	**HB**	**HBT**
Local variables						
EBITDA	Earnings before interest, taxes, depreciation, and amortization scaled by lagged total assets;	X		X	X	X
R&D	Research and development expenses divided by net sales, when missing reset to 0;	X		X	X	X
Advertising	Advertising expenses divided by net sales, when missing set to 0;	X		X	X	X
SG&A	Selling, general, and administrative expenses divided by net sales, when missing set to 0;	X		X	X	X

Capex	Reported capital expenditures divided by gross property, plant, and equipment;	X		X	X	X
Sales	The annual percentage change in net sales;	X		X	X	X
Leverage	The sum of long-term debt and long-term debt in current liabilities divided by total assets;	X		X	X	X
Cash	Cash and cash equivalents divided by total assets;	X		X	X	X
FOR	The firm has a non-missing, non-zero value for pre-tax income from foreign operations;	X		X	X	X
NOL	Net operating loss, an indicator if the firm has a non-missing value of tax loss carry-forward;	X		X	X	X
Size	The natural log of total assets;	X		X	X	X
Intangibles	The ratio of intangible assets to total assets;	X		X	X	X
PP&E	Gross property, plant, and equipment divided by total assets;	X		X	X	X
Network variables						
Closeness	Closeness centrality, the extent to which a firm is connected on average with all other firms;		X	X	X	X
Betweenness	Betweenness centrality, or how often a firm acts as a bridge between other firms in the network graph;		X	X		
Degree	Degree centrality, or the number of first (second for bipartite graphs) order neighbors;		X	X	X	X
PageRank	The importance of the firm in the network based on its neighbors and their importance, see also Page et al. [37]. The damping factor is set to 0.85 as suggested by Page et al.;		X	X	X	X
Lowdegree	The number of low-tax firms in the first (second for bipartite graphs) order neighborhood;		X	X	X	X
RLowdegree	Lowdegree relative to Degree;		X	X	X	X
WLowdegree	Weighted Lowdegree;		X	X		
Highdegree	The number of high-tax firms in the first (second for bipartite graphs) order neighborhood;		X	X	X	X
RHighdegree	Highdegree relative to Degree;		X	X	X	X
WHighdegree	Weighted Highdegree;		X	X		
AvgCETR	Average CETR value of first (second for bipartite graphs) order neighbors;		X	X	X	X
WAvgCETR	Weighted average CETR value of first (second for bipartite graphs) order neighbors. WAvgCETR cannot be calculated for the bipartite graphs since they contain only weights between board members and firms, and not firms mutually;		X	X		
MinCETR	Minimal CETR value of first (second for bipartite graphs) order neighbors;		X	X	X	X
MaxCETR	Maximal CETR value of first (second for bipartite graphs) order neighbors;		X	X	X	X
Sim	Number of first (second for bipartite graphs) order neighbors who are active in the same industry;		X	X	X	X
RSim	Number of first (second for bipartite graphs) order neighbors who are active in the same industry relative to Degree;		X	X	X	X
LowTri	Number of triangles with at least one low-tax firm;		X	X		
NLowTri	Number of triangles with no low-tax firms;		X	X		
RLowTri	Number of triangles with at least one low-tax firm relative to the total number of triangles;		X	X		
LowBM	Number of first order neighboring board members who are connected to at least two low-tax firms;				X	X

NLowBM	Number of first order neighboring board members who are connected to no low-tax firms;				X	X
CETRBM	Average CETR value of the firms the first order neighboring board members are connected to;				X	X
Busy	Average busyness of first order neighboring board members with busyness the number of firms the member is currently holding a board position. This variable was included based on Cashman et al. [9];				X	X
WLowBM	Weighted LowBM;					X
WNLowBM	Weighted NLowBM;					X
WCETRBM	Weighted CETRBM;					X
WBusy	Weighted Busy					X

3.5. Comparing model performance

We compare the predictive models in terms of their accuracy and their area under the ROC curve (AUC). Accuracy takes both the true positive (low-tax) and true negative (not low-tax) rate into account. Receiver operating characteristic (ROC) curves display the sensitivity (the true positive rate varying 0 to 1) versus the specificity (true negative rate varying 1 to 0). False positives are firms which are incorrectly classified as low-tax, and true positives are correctly classified as low-tax. As such, the closer the ROC curve is to the top left, and thus the higher the area under this curve, the better the model performs. AUC measures the probability that a randomly chosen low-tax firm gets a higher score than a randomly chosen not low-tax firm. Due to the relatively low number of low-tax firms compared to not low-tax firms, it is valuable to focus on AUC and sensitivity in stead of accuracy.

4. Results and discussion

4.1. Results

First, we train logistic regression models on the training data sets. All models were trained after feature selection was carried out on the training set

Table 3: Performance of the logistic regression models in terms of accuracy and AUC.

	2013		2014	
	Accuracy	**AUC**	**Accuracy**	**AUC**
Local model	88.89%	0.6627	89.61%	0.6882
Network unipartite model	88.97%	0.6015	89.69%	0.5519
Hybrid unipartite model	88.73%	0.6710	89.61%	0.6817
Hybrid unweighted bipartite model	90.09%	0.8399	90.01%	0.8287
Hybrid time-weighted bipartite model	89.53%	0.8394	89.29%	0.8332

leading to a selected subset of the variables. This feature selection process is based on the Akaike information criterion (AIC) measure and applied in a stepwise forward and backward manner. Afterwards, the remaining non-significant variables (p-value > 0.10) were consecutively omitted. This leads us to seven, seventeen, thirty, twenty-seven and thirty-one features for the local, network, hybrid unipartite, hybrid unweighted bipartite and hybrid time-weighted bipartite model respectively. As can be observed from Table 3, the hybrid unweighted bipartite model performs best in terms of AUC.

We furthermore note that it significantly outperforms the local, network and hybrid unipartite model (p-values < 0.0001 using the test of DeLong et al. [13]). The network model clearly performs worse, indicating the importance of including local variables. At the same time, we observe that the local model and the hybrid unipartite model perform similarly. These results indicate that network effects do play a significant and important role but they also illustrate the importance of a bipartite network which is able to extract more detailed features. For more details of the logistic regression models we refer to Appendix A.

Secondly, we train decision trees on the training data sets. For this

purpose we apply a conditional inference tree algorithm [24] and tune the parameter which must be exceeded in order to implement a split. This parameter is tuned by means of a four-fold cross-validation repeated ten times and consecutively set to 0.05. Moreover, we oversample the minority class (low-tax companies) in the training set to 20% of the sample size. Again, the hybrid bipartite models perform the best whereby the time-weighted model, with AUCs equal to 0.8144 and 0.8160 for 2013 and 2014 respectively, significantly outperforms the unweighted model, with resulting AUCs 0.7575 and 0.7558. Nevertheless, both models significantly outperform the local model (AUC = 0.6119 for 2013 and AUC = 0.6700 for 2014). However, the network unipartite and hybrid unipartite models are performing badly with AUCs equal to 0.5652 and 0.5832 for 2013, and AUCs of 0.5564 and 0.6414 for 2014.

Although these results do not improve the logistic regression models, we observe a benefit in modelling non-linear effects as the decision trees are still able to include the network effects, and correlations among the variables may exist. Therefore, we train random forests next using the algorithm of Breiman [6]. In order to determine the optimal value for the number of variables randomly sampled as candidates for each split, we apply a ten-fold cross-validation three times on the training set. We set the number of trees to an odd number in order to better be able to solve ties and an adequately high number relative to the number of variables included[1]. Table 4 shows that the hybrid bipartite models clearly outperform the other models in terms of AUC. The local and hybrid unipartite models perform worse but

[1]The exact number of trees depends on the number of variables and is minimally set to 1501.

Table 4: Performance of the random forest models in terms of accuracy and AUC.

	2013		2014	
	Accuracy	**AUC**	**Accuracy**	**AUC**
Local model	89.37%	0.7683	90.09%	0.7489
Network unipartite model	88.89%	0.6018	89.61%	0.5611
Hybrid unipartite model	89.05%	0.7496	89.77%	0.7474
Hybrid unweighted bipartite model	89.77%	0.8431	90.33%	0.8306
Hybrid time-weighted bipartite model	89.13%	0.8412	89.69%	0.8333

still surpass the network unipartite model. Figure 3 illustrates how the models compare to each other in terms of significant improvement in AUC. This comparison is also illustrated by means of ROC curves in Figures 4a and 4b for 2013 and 2014 respectively. Furthermore, as expected, all models show an improvement towards their logistic regression counterpart.

In order to provide a complete picture, we additionally trained random forests in a similar manner on the dataset from 2013 and tested the models for 2014. In comparison to the results for 2013 in Table 4, the same conclusions can be drawn. With an AUC of 0.8487, the hybrid unweighted bipartite models delivers the best results and outperforms the local model (with an AUC of 0.7512) and the hybrid unipartite model (with an AUC of 0.7483).

Next, we take a closer look at the sensitivity or ability of the model to identify low-tax firms, and specificity or ability of the model to identify firms which are not low-tax. Table 5 summarizes both metrics at a cut-off of 50% and an adapted cut-off so that the ratio of low-tax firms in the predictions equals the ratio of low-tax firms in the test sets. As such, the adapted cut-off will classify the 11% and 10%, for 2013 and 2014 respectively, most likely to

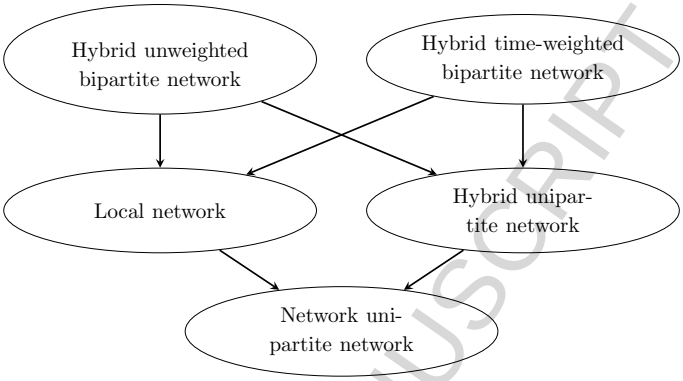
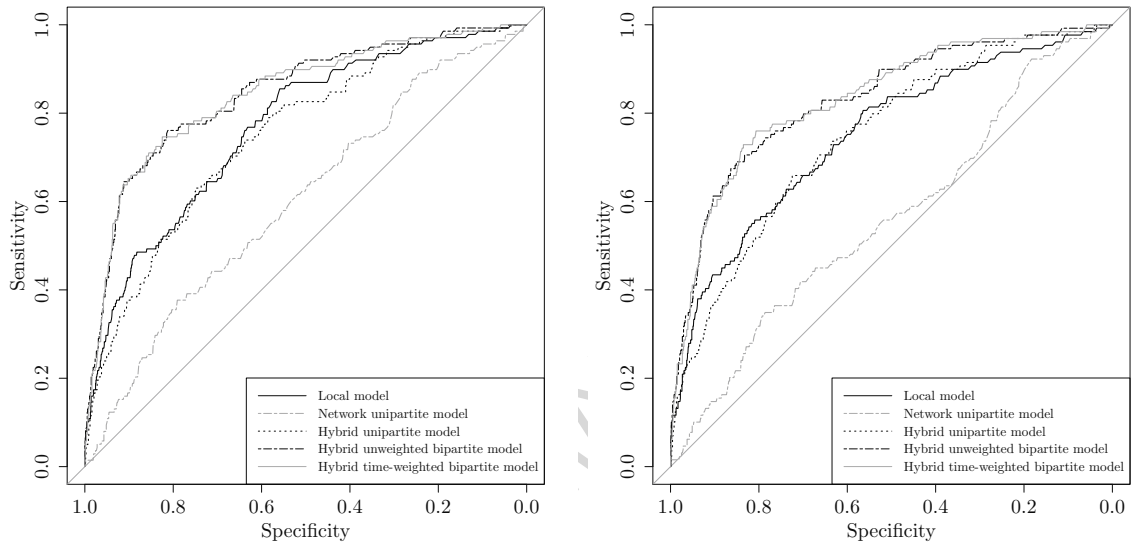


Figure 3: Domination graph [39] of random forest models based on pairwise comparison of AUC values [13]. Arrows indicate a significant performance improvement in AUC at a 0.1% significance level.

be low-tax firms as low-tax in fact. This metric will inform us whether we can correctly find all low-tax firms. We observe that the hybrid bipartite models are particularly better in identifying actual low-tax firms.

Finally, we shortly discuss the impact of applying sampling techniques before the random forest technique. Due to the fact that we work with unbalanced datasets —the ratio of low-tax firms is close to 10%—, it is interesting to research whether random undersampling of the majority group, i.e. not low-tax companies, improves the results. Undersampling to ratios 80/20 and 50/50 for respectively non-low-tax/low-tax firms does not significantly (p-values > 0.10) outperform the results of Table 4 in terms of AUC. Similarly, we applied SMOTE [10] with an oversampling and undersampling percentage of 200% and 400% respectively based on Chawla et al. [10], the ratio of the minority versus majority class and the original sample size. Also this sampling technique did not offer significant improvements.



(a) ROC curves for random forests validated for 2013 (b) ROC curves for random forests validated for 2014

Figure 4: ROC curves of the random forests validated for (a) 2013 and (b) 2014 representing the local, network unipartite, hybrid unipartite, hybrid unweighted bipartite and hybrid time-weighted model.

4.2. Discussion

We have created tax avoidance prediction models using three popular machine learning techniques, namely logistic regression, decision trees and random forests. All techniques strongly indicate the potential of including characteristics extracted from a network where firms are linked if they share board members. Moreover, we note that (1) network variables cannot replace firm characteristics for tax avoidance prediction but complement them; and (2) that including bipartite network characteristics which are more detailed with regards to the board members themselves provides us with important information. We also remark that weighing the edges in the bipartite network by the membership of the board member in time, does not improve performance.

Table 5: The sensitivity (sens) and specificity (spec) of the random forest models for 2013 and 2014. Both metrics are calculated for a 50% cut-off rate (50) and an adapted cut-off rate (ad) similar to the actual ratio of low-tax firms in the test sets.

	2013		**2014**	
	Sens 50	**Spec 50**	**Sens 50**	**Spec 50**
Local model	0.04348	0.9991	0.04651	0.9991
Network unipartite model	0.007246	0.9982	0.007752	0.9982
Hybrid unipartite model	0.04348	0.9955	0.04651	0.9955
Hybrid unweighted bipartite model	0.2101	0.9829	0.2171	0.9822
Hybrid time-weighted bipartite model	0.2826	0.9668	0.2946	0.9661
	Sens ad	**Spec ad**	**Sens ad**	**Spec ad**
Local model	0.3768	0.9227	0.3876	0.9296
Network unipartite model	0.1739	0.8976	0.1473	0.9020
Hybrid unipartite model	0.3406	0.9182	0.3101	0.9207
Hybrid unweighted bipartite model	0.4928	0.9371	0.4496	0.9367
Hybrid time-weighted bipartite model	0.5000	0.9380	0.4574	0.9376

Next, we take a closer look at the variables of the hybrid unweighted bipartite network. First, we take a look at the variables included in the logistic regression model. Their details are noted in Appendix A and visualized by means of a colored nomogram in Figure 5. We observe that there are three important characteristics of a firm: a lower EBITDA (p-value < 0.05), a non-missing value for its tax loss carry-forward (p-value < 0.05) and a higher PP&E (p-value < 0.001) lead to an increased probability of being a low-tax firm. For the network characteristics, a higher number of neighboring low-tax firms (p-value < 0.05), a higher average CETR of a firm's neighbors (p-value < 0.01), a lower number of board members who are not connected to low-tax firms (p-value < 0.001), and a lower average CETR of the neighbors of a firm's board members (p-value < 0.001), lead to a higher probability of being a low-tax firm. The direction of the AvgCETR estimate seems unexpected but might be due to interaction effects not cap-

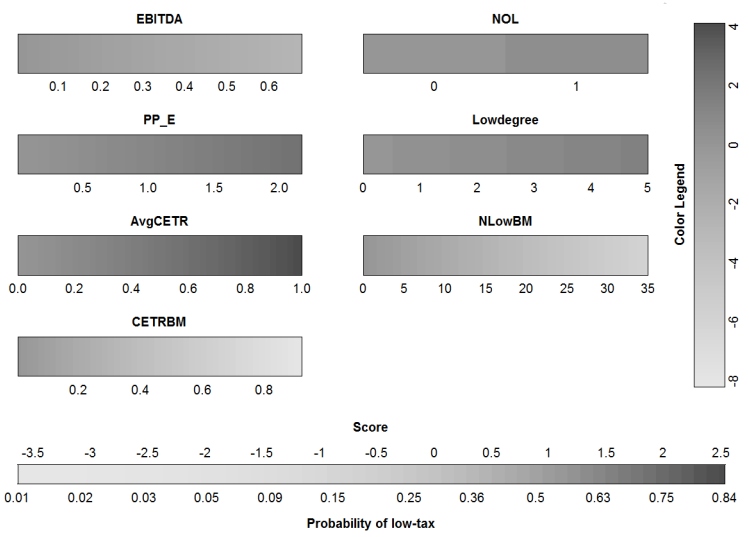


Figure 5: Colored nomogram. The color indicates the extent to which a variable contributes to the probability of being a low-tax firm, and can be converted to points by means of the Color Legend (on the right). To calculate the final probability, all points can be summed and converted by means of the Score bar (at the bottom). This visualization was created based on the work of Van Belle & Van Calster [42].

tured by the logistic regression model. When we, in addition, take a closer at the hybrid unipartite model, we observe a positive effect of betweenness (p-value < 0.01). This variable can be interpreted as the information which flows through this company via the board members. The higher the betweenness, the better a firm is able to control this information flow [34]. This increases support for the idea of a valuable information flow on tax strategies between firms through their board members.

We can furthermore derive the importance of the specific local and network variables in the random forest model by studying decreases in node impurity measured by the Gini index if we would remove a particular variable from the decision trees, see Figure 6. We notice that two bipartite network features receive a high importance for the creation of the random

forest, namely if firms have board members who are not involved in low-tax firms and the average CETR of the firms a particular firm's board member is connected to. This is consistent with the previously mentioned idea that the knowledge of board members in the network is crucial. Next, three local variables rank high, the PP&E, EBITDA and Sales. Clearly, both firm as well as network characteristics play an important role in the creation of our best performing random forest model. The reader is referred to Appendix B for more details on the variable importance in this model.

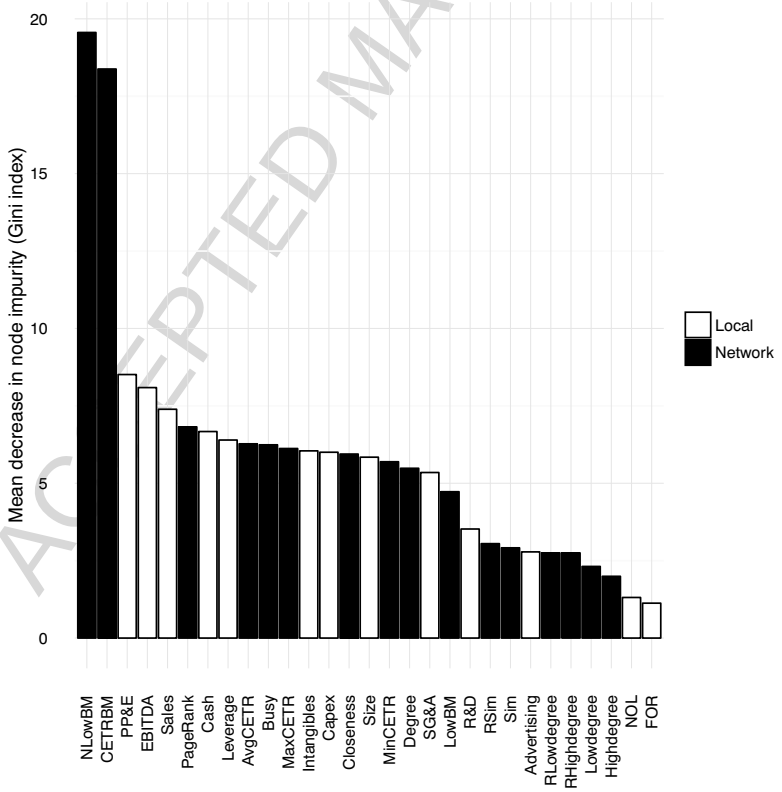


Figure 6: Mean decrease in node impurity measured by the Gini index if a particular variable is removed from the variable set.

Lastly, we want to discuss the related topic of tax evasion. The tax

avoidance measure merely picks up the firm's ability to, by means of tax planning, pay a low amount of taxes relative to its earnings over an extended period of time [15]. We do not infer that these firms realize these tax rates as a result of illegal activities. For tax evasion, different proxies exist that can be explored in future studies. Yet, a firm who consistently pays low taxes might be at risk due to taking on a more aggressive tax policy. Therefore, our models might still inform intermediaries (e.g. financial analysts) who either assess the firm's risk, or tax authorities who want to target firms for further investigation.

4.3. Further research

This paper clearly demonstrates the potential of social network characteristics for tax avoidance prediction. Nevertheless, further research could be undertaken. For example, we observe that time-weighted edges do not enhance the bipartite network. However, this does not necessarily reduce its potential given that good results were previously obtained in the fraud detection domain [44]. Depending on the dataset and resulting network, the decay factor γ, see Equation 2, could be further fine-tuned or different weights could be assigned to the edges to examine the application of information which diminishes over time. These weights could for example take job characteristics of the board member into account. Next, the social network could be created with the board members as a starting point instead of the firms. In this sense, social ties between board members could even be taken into account. Finally, it could be interesting to research whether different pre-processing or machine learning techniques are able to further improve the performance, e.g. artificial neural networks, support vector machines, etc. Alternatively, regression techniques could be applied to make

numerical predictions of tax avoidance. These predictions can then be used as an input to the classification of low-tax firms.

Furthermore, our results hint at the fact that firms may tap into different networks. Namely, firms with directors who have few links to low-tax firms, may have more difficulties in maintaining a low-tax policy because either they miss the knowledge, or because they prefer to have directors who remain critical against tax avoidance. Yet, connections to low-tax firms may signal risk as directors may be hired for their expertise to develop and maintain a low-tax strategy. Future research could investigate the profiles of each of these directors in more detail to gain deeper insights on what drives the spill-over effects across companies in the network.

5. Conclusion

In this paper, we developed a tax avoidance prediction model which incorporates network characteristics of firms. This network was constructed based on shared board members. Consecutively, three analytics techniques, logistic regression, decision trees and random forests; were applied on firm-specific characteristics, on an elaborate set of network characteristics and on different combinations of both. Hereby, unipartite network characteristics which only include network details about the firms, as well as bipartite network characteristics which also include network details about board members, were examined. Our hybrid bipartite random forest model performed best with an 7 pp increase in AUC compared to its local counterpart. As such, we are able to better predict which firms are low-tax and which are not. To the best of our knowledge, we are the first to apply and validate predictive analytics models that include a large spectrum of network fea-

tures to the domain of financial reporting and tax avoidance. In doing so, we offer new insights that can assist companies in their tax planning and their search for attracting the right expertise for their boards. The idea that board members who have previously seated in low-tax firms are conveying their knowledge and expertise, is further motivated by our findings. Firms who lack connections to low-tax firms and the knowledge (by having many board members not connected to low-tax firms) are less likely to be classified as low-tax. Furthermore, because we achieved increased predictive power by including network features, regulatory agencies and financial analysts also benefit from our models. The ability to better predict future low-tax firms may help tax authorities in generating more revenue by being better able to identify companies that potentially engage in more aggressive tax policies and thus deserve further investigation.

Funding

The work performed by Sander De Groote was supported by the Research Foundation − Flanders (FWO).

References

[1] Armstrong, C. S., Blouin, J. L., Jagolinzer, A. D., & Larcker, D. F. (2015). Corporate governance, incentives, and tax avoidance. *Journal of Accounting and Economics*, *60*, 1–17.

[2] Armstrong, C. S., Blouin, J. L., & Larcker, D. F. (2012). The incentives for tax planning. *Journal of Accounting and Economics*, *53*, 391–411.

[3] Baesens, B., Van Vlasselaer, V., & Verbeke, W. (2015). Social network analysis for fraud detection. In *Fraud Analytics: Using Descriptive, Predictive, and Social Network Techniques* (pp. 207–278). Hoboken, NJ: Wiley.

[4] Bianchi, P. A., Falsetta, D., Minutti-Meza, M., & Weisbrod, E. H. (2016). Professional networks and client tax avoidance: Evidence from the Italian statutory audit regime. Available at SSRN: `https://ssrn.com/abstract=2601570` (Last revised: 03/16/2016).

[5] Bizjak, J., Lemmon, M., & Whitby, R. (2009). Option backdating and board interlocks. *Review of Financial Studies*, *22*, 4821–4847.

[6] Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

[7] Brown, J. L., & Drake, K. D. (2014). Network ties among low-tax firms. *The Accounting Review*, *89*, 483–510.

[8] Bruynseels, L., & Cardinaels, E. (2014). The audit committee: Management watchdog or personal friend of the CEO? *The Accounting Review*, *89*, 113–145.

[9] Cashman, G. D., Gillan, S. L., & Jun, C. (2012). Going overboard? On busy directors and firm value. *Journal of Banking & Finance*, *36*, 3248–3259.

[10] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

[11] Chiu, P.-C., Teoh, S. H., & Tian, F. (2013). Board interlocks and earnings management contagion. *The Accounting Review*, *88*, 915–944.

[12] Dechow, P. M., & Tan, S. T. (2016). How do accounting practices spread? An examination of law firm networks and stock option backdating. Available at SSRN: `https://ssrn.com/abstract=2688434` (Last revised: 02/24/2016).

[13] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837–845.

[14] Desai, M. A., & Dharmapala, D. (2006). Corporate tax avoidance and high-powered incentives. *Journal of Financial Economics*, *79*, 145–179.

[15] Dyreng, S. D., Hanlon, M., & Maydew, E. L. (2008). Long-run corporate tax avoidance. *The Accounting Review*, *83*, 61–82.

[16] Dyreng, S. D., Hanlon, M., & Maydew, E. L. (2010). The effects of executives on corporate tax avoidance. *The Accounting Review*, *85*, 1163–1189.

[17] Gallemore, J., & Labro, E. (2015). The importance of the internal information environment for tax avoidance. *Journal of Accounting and Economics*, *60*, 149–167.

[18] Graham, J. R., Hanlon, M., Shevlin, T., & Shroff, N. (2014). Incentives for tax

planning and avoidance: Evidence from the field. *The Accounting Review*, *89*, 991–1023.

[19] Han, J., Bose, I., Hu, N., Qi, B., & Tian, G. (2015). Does director interlock impact corporate R&D investment? *Decision Support Systems*, *71*, 28–36.

[20] Hanlon, M., & Heitzman, S. (2010). A review of tax research. *Journal of Accounting and Economics*, *50*, 127–178.

[21] Hanlon, M., & Slemrod, J. (2009). What does tax aggressiveness signal? evidence from stock price reactions to news about tax shelter involvement. *Journal of Public Economics*, *93*, 126–141.

[22] Hasan, M. A., & Zaki, M. J. (2011). A survey of link prediction in social networks. In C. C. Aggarwal (Ed.), *Social Network Data Analytics* (pp. 243–275). Boston, MA: Springer US.

[23] Horton, J., Millo, Y., & Serafeim, G. (2012). Resources or power? Implications of social networks on compensation and firm performance. *Journal of Business Finance & Accounting*, *39*, 399–426.

[24] Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*, 651–674.

[25] Jiang, C., Kubick, T. R., Miletkov, M. K., & Wintoki, M. B. (forthcoming). Offshore expertise for onshore companies: Director connections to island tax havens and corporate tax policy. *Management Science*, *0*.

[26] Kramer, S., Lavrač, N., & Flach, P. (2001). Propositionalization approaches to relational data mining. In S. Džeroski, & N. Lavrač (Eds.), *Relational Data Mining* (pp. 262–291). Berlin, Heidelberg: Springer.

[27] Lanis, R., & Richardson, G. (2011). The effect of board of director composition on corporate tax aggressiveness. *Journal of Accounting and Public Policy*, *30*, 50–70.

[28] Larcker, D. F., So, E. C., & Wang, C. C. (2013). Boardroom centrality and firm performance. *Journal of Accounting and Economics*, *55*, 225–250.

[29] Lessmann, S., Baesens, B., Seow, H., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*, 124–136.

[30] Ma, Z., Sheng, O. R., & Pant, G. (2009). Discovering company revenue relations from news: A network approach. *Decision Support Systems*, *47*, 408–414.

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

[32] McGuire, S. T., Omer, T. C., & Wang, D. (2012). Tax avoidance: Does tax-specific industry expertise make a difference? *The Accounting Review*, *87*, 975–1003.

[33] Minnick, K., & Noga, T. (2010). Do corporate governance characteristics influence tax management? *Journal of Corporate Finance*, *16*, 703–718.

[34] Neuman, S. S. (2014). Effective tax strategies: It's not just minimization. doi:10.2139/ssrn.2496994.

[35] Omer, T. C., Shelley, M. K., & Tice, F. M. (2014). Do well-connected directors affect firm value? *Journal of Applied Finance*, *24*, 17–32.

[36] Omer, T. C., Shelley, M. K., & Tice, F. M. (2016). Do director networks matter for financial reporting quality? Evidence from restatements. doi:10.2139/ssrn.2379151.

[37] Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The PageRank citation ranking: bringing order to the web*. Technical Report 1999-66 Stanford InfoLab.

[38] Rego, S. O., & Wilson, R. (2012). Equity risk incentives and corporate tax aggressiveness. *Journal of Accounting Research*, *50*, 775–810.

[39] Rossetti, M., Stella, F., & Zanker, M. (2016). Contrasting offline and online results when evaluating recommendation algorithms. In *Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, September 15-19, 2016* (pp. 31–34).

[40] Schabus, M. (2016). Do director networks help managers plan better? Available at SSRN: `https://ssrn.com/abstract=2824070` (Last revised: 12/06/2016).

[41] Slemrod, J. B. (2016). *Tax Compliance and Enforcement: New Research and Its Policy Implications*. Technical Report 1302 Ross School of Business. Available at SSRN: `https://ssrn.com/abstract=2726077` (Last revised: 02/17/2016).

[42] Van Belle, V., & Van Calster, B. (2015). Visualizing risk prediction models. *PLoS ONE*, *10*, e0132614.

[43] Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2015). APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems*, *75*, 38–48.

[44] Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2017). GOTCHA! Network-based fraud detection for social security fraud. *Management*

Science, *63*, 3090–3110.

Appendix A. Logistic regression

Table A.1 indicates for each model the estimates of each variable and whether it is significant. Note that not every variable is relevant for each model, see Table 2, and that some variables were excluded after feature selection.

Table A.1: For each variable it is depicted whether the model includes the variable after feature selection and, if included, it shows the estimated effect and the significance of the effect.

Variables	**Local model**	**Network unipartite model**	**Hybrid unipartite model**	**Hybrid unweighted bipartite model**	**Hybrid time-weighted bipartite model**
Firm characteristics					
Intercept	−2.1261****	−2.3654****	−0.8191	−0.8920*	−0.8549*
EBITDA	−5.2956****		−5.7446****	−3.9855**	−3.8716**
R&D	5.8106**		5.7119**	Not included	Not included
Advertising	Not included		Not included	Not included	Not included
SG&A	−1.7326*		−2.0908*	Not included	Not included
Capex	Not included		Not included	Not included	Not included
Sales	Not included		Not included	Not included	Not included
Leverage	0.9099*		1.0778**	Not included	Not included
Cash	Not included		Not included	Not included	Not included
FOR	−0.6063**		−0.5622**	Not included	Not included
NOL	0.7309***		0.7716***	0.6272**	0.6282**
Size	Not included		−0.1684**	Not included	Not included
Intangibles	Not included		Not included	Not included	Not included
PP&E	1.0486****		1.0151****	1.0917****	1.1315****
Network characteristics					
Closeness		Not included	Not included	Not included	Not included
Betweenness		340.0107***	211.1058***		

Degree		−200.0625**	Not included	Not included	Not included
PageRank		Not included	Not included	Not included	Not included
Lowdegree		Not included	Not included	0.2965**	0.2835**
RLowdegree		Not included	Not included	Not included	Not included
WLowdegree		Not included	Not included		
Highdegree		−0.3115**	−0.3115**	Not included	Not included
RHighdegree		Not included	Not included	Not included	Not included
WHighdegree		Not included	Not included		
AvgCETR		Not included	Not included	4.0542***	3.8103***
WAvgCETR		Not included	Not included		
MinCETR		Not included	Not included	Not included	Not included
MaxCETR		1.1637*	Not included	Not included	Not included
Sim		Not included	Not included	Not included	Not included
RSim		Not included	Not included	Not included	Not included
LowTri		Not included	Not included		
NLowTri		Not included	Not included		
RLowTri		Not included	Not included		
LowBM[2]				Not included	Not included
NLowBM				−0.1704****	−0.1863****
CETRBM				−8.8259****	Not included
Busy				Not included	Not included
WLowBM					Not included
WNLowBM					Not included
WCETRBM					−10.6228****
WBusy					Not included

*p-value < 0.1; **p-value < 0.05; ***p-value < 0.01; ****p-value < 0.001

[2]LowBM was excluded after feature selection presumably because of its correlation to NLowBM. The more directors who are connected to non-low tax firms (NLowBM), the less likely that there are directors connected to two or more low tax firms (LowBM). Exchanging NLowBM for LowBM shows that this variable is positive and significant at a 5% significance level in the hybrid unweighted bipartite model and at a 1% in the hybrid time-weighted bipartite model. Having board members with at least two connections to low-tax firms thus increases the probability of being low-tax.

Appendix B. Variable importance in the hybrid unweighted bipartite random forest model

To interpret which variables are the most important in a random forest model, we can study the mean decrease in node impurity, in terms of Gini index, and the mean decrease in accuracy if we would leave out this variable during the construction of the decision trees. The details can be observed in Table B.1.

Table B.1: Mean decrease in node impurity and accuracy of each variable if it would not have been included in the decision trees of the hybrid unweighted bipartite random forest. Network characteristics are emphasized in bold in the first column.

Variables	**Mean decrease in node impurity**	**Mean decrease in accuracy**
NLowBM	***19.5603***	***0.01869***
CETRBM	***18.3801***	***0.01138***
PP&E	8.5114	0.002731
EBITDA	8.0896	***0.002485***
Sales	7.3901	0.001047
PageRank	6.8208	0.003354
Cash	6.6708	0.001160
Leverage	6.3962	0.0009347
AvgCETR	6.2739	0.003184
Busy	6.2419	0.0007780
MaxCETR	6.1224	0.003050
Intangibles	6.0466	0.001789
Capex	6.0018	0.0008163
Closeness	5.9434	0.003846
Size	5.8421	0.001487
MinCETR	5.6951	0.002137
Degree	5.4840	0.003788
SG&A	5.3471	0.001520
LowBM	***4.7263***	***0.002891***
R&D	3.5224	0.001600

RSim	3.0493	0.0009655
Sim	2.9148	0.0009793
Advertising	2.7832	0.0002154
RLowdegree	2.7507	0.0009238
RHighdegree	2.7507	0.001378
Lowdegree	2.3139	0.001413
Highdegree	1.9948	0.001268
NOL	1.3112	*0.0005218*
FOR	1.1271	0.0003681

p-value $<$ *0.1* (in italic); ***p-value*** $<$ ***0.05*** (in italic, bold); ***p-value*** $<$ ***0.01*** (in italic, bold, underlined)

Biography

Jasmien Lismont obtained her PhD in Business Economics in 2018 at the Faculty of Economics and Business, KU Leuven, Belgium. At KU Leuven, she previously also received a Master's degree in Information Systems Engineering (magna cum laude) in 2014 after which she started her research at the Department of Decision Sciences and Information Management. Her main topics of interest include the strategic impact of analytics and big data on business; marketing analytics; social network analytics; direct marketing; and corporate governance of analytics. In addition, she has presented and participated at various international conferences.

Eddy Cardinaels is full Professor of accounting at KU Leuven and part-time professor at Tilburg University. His work focuses on experimental research in accounting, studying the conditions under which decision makers can benefit from recent accounting innovations such as ABC, BSC, and time driven costing. Other work explores how social motives, trust between superiors, subordinates and/or business partners affect honest reporting in business settings, and how such trust affects inter-firm negotiations. A second line of research focuses on archival research in accounting examining the effects of social networks in publicly listed companies and the impact of corporate governance quality and incentive mechanisms for both listed companies and private non-for-profit firms. He has published in the top accounting journals and has served as editor for The Accounting Review (2014-2017) and as Guest Editor for the special issue of European Accounting Review 'Accounting Insights from Health Care'. He further is an editorial board member for Contemporary Accounting Research and Journal of Auditing Practice and Theory.

Liesbeth Bruynseels is associate professor at the department of Accounting, Finance and Insurance at KU Leuven, Belgium. Her research focuses on corporate governance; board dynamics; social connections; auditor judgment and decision-making and determinants of audit quality. Her work has been published in leading accounting journals such as The Accounting Review; Accounting, Organizations and Society and Auditing: A Journal of Practice and Theory. She also acts as associate editor at The European Accounting Review, where she handles papers related to auditing and corporate governance.

Sander De Groote is a PhD candidate at the department of Accounting, Finance and Insurance at KU Leuven, Belgium. In 2015, Sander obtained a FWO-research grant fully funding his PhD project which he will be finishing in 2018. His research interests include corporate governance, board dynamics, director insider trading and director compensation.

Bart Baesens is a professor at KU Leuven, Belgium, and a lecturer at the University of Southampton, United Kingdom. He has done extensive research on predictive analytics, data mining, web analytics, fraud detection, and credit risk management. His findings have been published in well-known international journals and presented at international top conferences. He is also author of the book Analytics in a Big Data World, published in 2014; and co-author of the books Credit Risk Analytics, published in 2016; Fraud Analytics using Descriptive, Predictive & Social Network Techniques, published in 2015; and Credit Risk Management, published in 2008.

Wilfried Lemahieu obtained a PhD from the Faculty of Economics and Business (FEB) at KU Leuven, Belgium. At present, he holds a position as full professor at the Department of Decision Sciences and Information Management of that faculty. His teaching includes Database Management, Enterprise Information Management and Management Informatics. His current research focuses on big data storage

and integration, data quality, business process management and service oriented architectures. His research was published in several international journals and he is a frequent lecturer for both academic and industry audiences. Since 2017, he assumes the position of Dean at FEB.

Jan Vanthienen is full professor of information systems at KU Leuven, Department of Decision Sciences and Information Management. He has built extensive expertise in the areas of decision modeling and analytics, business rules & processes, and information management. He has published extensively in top international journals (such as Machine Learning, Management Science, Journal of Machine Learning Research, MIS Quarterly, IEEE Transactions on Neural Networks), and conferences (ICIS, CAiSE, BPM, …). He received an IBM Faculty Award on smart decisions, and the Belgian Francqui Chair at FUNDP. He is co-founder and president-elect of the Benelux Association for Information Systems (BENAIS).

Highlights

- First tax avoidance prediction model by means of advanced social network analytics
- Network of firms which are connected if they share board members now or in the past
- A 7 pp increase in AUC thanks to including uni- and bipartite network variables
- Tax rate of neighboring firms and number of low-tax neighbors has an influence
- Implies carry-over of knowledge between firms by board members