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 complement previous literature by being the first to build a predictive model including a larger variation of network features. We construct a network of firms which are connected through shared board membership. Three analytical techniques are applied creating five models using either firm characteristics or network characteristics or different types of combinations of both. A random forest which includes firm characteristics, network characteristics of firms and network characteristics of board members provides the best performance with an increase of 7% 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 likelihood of being a low-tax firm. Simi-

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

larly, 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. Our results are informative for companies as to the director expertise they want to attract. Additionally, regulatory agencies can use our insights to predict which firms are likely to be low-tax companies and thus require further investigation.
Keywords: analytics, tax avoidance, social network analytics, board interlocks, predictive analytics

---

1. Introduction

There is considerable variation in taxes being paid among corporate organizations (Hanlon & Heitzman, 2010; Christensen et al., 2015, p. 1919). While firms enjoy benefits of tax avoidance by lower taxes being paid, tax planning does not come without risk as tax authorities may impose fines and penalties for tax evasion, and tax avoidance may involve significant political and reputational costs (Lanis & Richardson, 2011). Motivated by the variation and the different trade-offs, researchers start to explain why firms engage in tax avoidance. Many studies focus on firm-specific variables to explain tax avoidance and the various incentives that directors receive (Gupta & Newberry, 1997; Rego, 2003; Desai & Dharmapala, 2006; Minnick & Noga, 2010; Armstrong et al., 2012). The literature also includes different governance variables, looks at the quality of information systems, and incorporates the various types of expertise in the board or the audit office, indicating that tax planning does require a certain level of expertise and knowledge (Lanis & Richardson, 2011; McGuire et al., 2012; Robinson et al., 2012; Gallemore & Labro, 2015). This paper uses techniques from the

social network analytics domain to develop a predictive model for tax avoidance. Motivated from the idea that a certain level of expertise is required for tax avoidance, we look at how firms are connected through shared directorships and how shared knowledge in the network and connections to low-tax firms (through director sharing) can be informative for tax avoidance. We create a predictive model using firm-specific variables that prior literature has typically incorporated complemented by network features (Page et al., 1998; Baesens et al., 2015; Van Vlasselaer et al., 2017). We found that a combination of firm characteristics and network characteristics provides the best predictive performance. As such, a hybrid model combining both types of characteristics is able to identify more low-tax firms.

First, we discuss related research on tax avoidance and social network analytics to illustrate the importance and novelty of our study in Section 2. 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 illustrated that human actors in firms have access to specific human capital and that such knowledge of corporate directors seems to travel across a director's network. Bizjak et al. (2009) show, for example, that firms who have a board member of a firm previously identified as a backdating firm, are more likely to backdate stock options themselves. In the same context, Dechow & Tan (2016) discovered that backdating firms are more highly connected via shared law firms. Horton et al. (2012); Larcker et al. (2013); and Omer et al. (2014) take a closer look at firm performance

and how directors' connectedness impacts this. Schabus (2016) 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. (2013) indicate that earnings management contagion occurs more often for firms who have directors in common.

Following this line of reasoning, researchers start to look at the impact of network effects on tax avoidance. This network consists of either companies or directors that are linked or connected. 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, because they share their job title, or because they know each other in a social context (Bruynseels & Cardinaels, 2014; Omer et al., 2016). Dyreng et al. (2010) examine, for example, whether executive effects, next to firm characteristics, impact tax avoidance. Tracking individual executives across companies, they found 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 it could be interesting to include network effects besides firm characteristics. Bianchi et al. (2016) look at auditor ties and found that better connected auditors have an impact on their clients' tax avoidance. Neuman (2014) includes directors' connections in order to gain insight into firms' tax planning. For this purpose, Neuman extracts four centrality features from a social network of directors, namely degree, betweenness, closeness and eigenvector centrality. Brown & Drake (2014) examine the impact of board interlocks on tax avoidance rates by extracting the number of ties to low-tax firms. They found that firms who have more

board members tied to low-tax firms, enjoy lower tax rates themselves.

However, all previous research investigating the effect of social networks for tax avoidance do this in a descriptive manner. They focus on how well one firm is connected to other firms via shared directorship. We supplement this literature by first of all developing a more extensive set of network measures which are validated by means of advanced machine learning techniques, in order to offer a broader picture of which network features are more informative for tax planning activities of firms. We complement prior literature by creating a predictive model for tax avoidance and thus providing insights on the predictive value of some of the network features, relative to firm-specific variables. This allows us to speak about the economic importance of network effects in the tax planning of firms and to the literature that tries to validate the predictive value of different social network techniques (Hasan & Zaki, 2011, p. 246; Baesens et al., 2015).

We provide insights to apply appropriate methods and techniques for the creation of a predictive model for tax avoidance. Such predictive models are of interest to management, shareholders, and directors that often are involved in tax planning strategies for the company (Graham et al., 2014). Shareholders can benefit from a low tax rate. Companies rank increased earnings per share as one of the key reasons for engaging in tax planning activities (Graham et al., 2014). Also management and corporate directors (including tax directors) often receive significant financial incentives which further may increase the motivation to engage in tax avoidance (Slemrod, 2004; Armstrong et al., 2012). 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 di-

rectors. Second, albeit different but maybe even more important, our models might also inform intermediaries (e.g. financial analysts) who either assess the firm's risk or tax authorities which want to target firms for investigation. Aggressive tax avoidance also raises risks for investors of the companies, as companies may become under higher public scrutiny. Financial analysts can incorporate this risk better, based on the parameters we predict to be crucial for tax avoidance. Additionally, as noted by Slemrod (2016), US regulators increased their focus on tax evasion after the financial crisis of 2008 both in terms of policy and enforcement. The Internal Revenue Service (IRS), given limited budgets, also uses modern data analysis techniques to identify potential tax evaders. 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 likely to be at risk.

3. Materials and methods

3.1. Data description

We have collected firm characteristics data from Compustat and data on corporate board members from BoardEx for fiscal years 2004 until 2014. 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 (2014), 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 of the reasons listed by Neuman (2014). She claims that CETR is a more representative and

comprehensive measure of a firm's tax planning strategy.

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

Next, we identify low-tax firms as firms ranked in the lowest quintile based on CETR and adjusted for industry mean (Brown & Drake, 2014). Similarly, high-tax firms are distinguished. We specifically focus on categorization instead of a continuous tax rate since corporate governance effects are stronger for more extreme formats of tax avoidance (Armstrong et al., 2015). We start with defining local variables, only based on firm characteristics, and explain the extracted network variables in the next section. Local variables are based on the definitions of Dyreng et al. (2010) and can be found in Table 1.

3.2. Building a social network

A graph or network consists of nodes, also referred to as vertices, and edges or the links that connect nodes. We create three types of graphs. (1) First, we create a unipartite, undirected, weighted graph where nodes are the firms which are connected if they have current or previous board members in common. We weigh each edge by the number of shared board members. This graph is undirected because the edges have no arrow and they do not flow from one firm to another; and it is unipartite because it only has one type of nodes, namely firms. (2) Secondly, we create a bipartite, undirected, unweighted graph, illustrated in Figure 1. A bipartite graph has two types of nodes, which are in this case firms and board members. Each firm is connected with board members and each board member is connected with one or more firms. Each edge now has the same weight of 1, leading to an unweighted graph. (3) Thirdly, we create a bipartite,

Figure 1: Bipartite graph with firms (black nodes) and board members (grey nodes) which are connected by board membership. The larger the firm node, the lower its CETR rate. The visualization was created by means of the OpenOrd (Martin et al., 2011) algorithm which uses simulated annealing to cluster nodes.

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 the connection with this firm. 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. (2017) with a decay factor γ set to 0.6 and h representing the number of years the board member is not sitting on the

board anymore with $h = 0$ for current board members. The decay factor was determined based on the time frame of our training dataset running from fiscal years 2004 until 2012.

$$\begin{aligned} &\mathrm{W}_{i,j} = e^{-\gamma h} && \text{if a relationship exists between firm } i \text{ and board member } j \\ &\mathrm{W}_{i,j} = 0 && \text{otherwise} \end{aligned} \tag{2}$$

There are multiple ways to use network characteristics in an analytical model (Macskassy & Provost, 2007; Verbeke et al., 2014). We chose to extract features from the network so that we are able to use them by non-relational predictive analytics techniques. Moreover, this technique allows us to analyze the effects of the network features. This process is also referred to as featurization or propositionalization (Kramer et al., 2001). Table 1 presents the features we deduced from the 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 firms 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. (2017) in a fraud detection context. A triangle 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 Table 1. Note that the

betweenness was not calculated for the nodes in the bipartite graphs due to the large computation efforts for this measure.

Table 1: 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 (see Section 3.3).

Variable	**Description**	**L**	**N**	**HU**	**HB**	**HBT**
Firm characteristics						
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 characteristics						
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. (1998). 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;		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. (2012);				X	X
WLowBM	Weighted LowBM;					X
WNLowBM	Weighted NLowBM;					X
WCETRBM	Weighted CETRBM;					X
WBusy	Weighted Busy					X

3.3. Methodology

By means of predictive analytical models, we aim to classify a firm as low-tax or not. For this purpose we train our models on a training set covering 1,032 firms from fiscal years 2004 until 2012. This means that we take the firm characteristics of 2011 and the tax avoidance rate of 2012. Furthermore, the network is created using the board membership data of 42,298 directors from 2004 until 2011. Of the training dataset 9.11% are low-tax firms while 17.83% are high tax firms. Next, we compare the performance of our models on two out-of-time validation 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. Also for the validation 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, medium or high) and from 2013 to 2014 7.07% changed. This training, testing and validation process is depicted in Figure 2.

We compare the predictive models in terms of accuracy and 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 versus the specificity. Frequently, this is also represented as the true positive rate, versus the the true negative rate or 1 minus the false positive rate. 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. Thus, the AUC measures the probability that a randomly chosen low-tax firm gets a higher score than a randomly chosen not low-tax firm.

We apply three techniques, namely logistic regression, decision trees and random forests. Logistic regression and decision trees are common techniques for classification tasks. The decision tree algorithm applied follows Breiman et al. (1984) quite closely (Therneau et al., 2015). Random forests are an ensemble technique which constructs multiple decision trees and combines them into one model. For this purpose, we use the random forest algorithm of Breiman (2001). We explicitly include this technique because various benchmarking studies illustrated its superior performance, e.g. Lessmann et al. (2015).

Next, we create five models with each technique. The specific variables included in each model are depicted in Table 1. 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

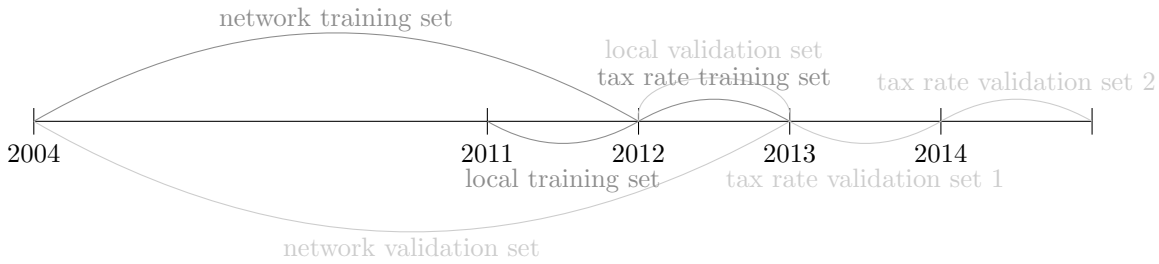


Figure 2: Data collection for training and validation.

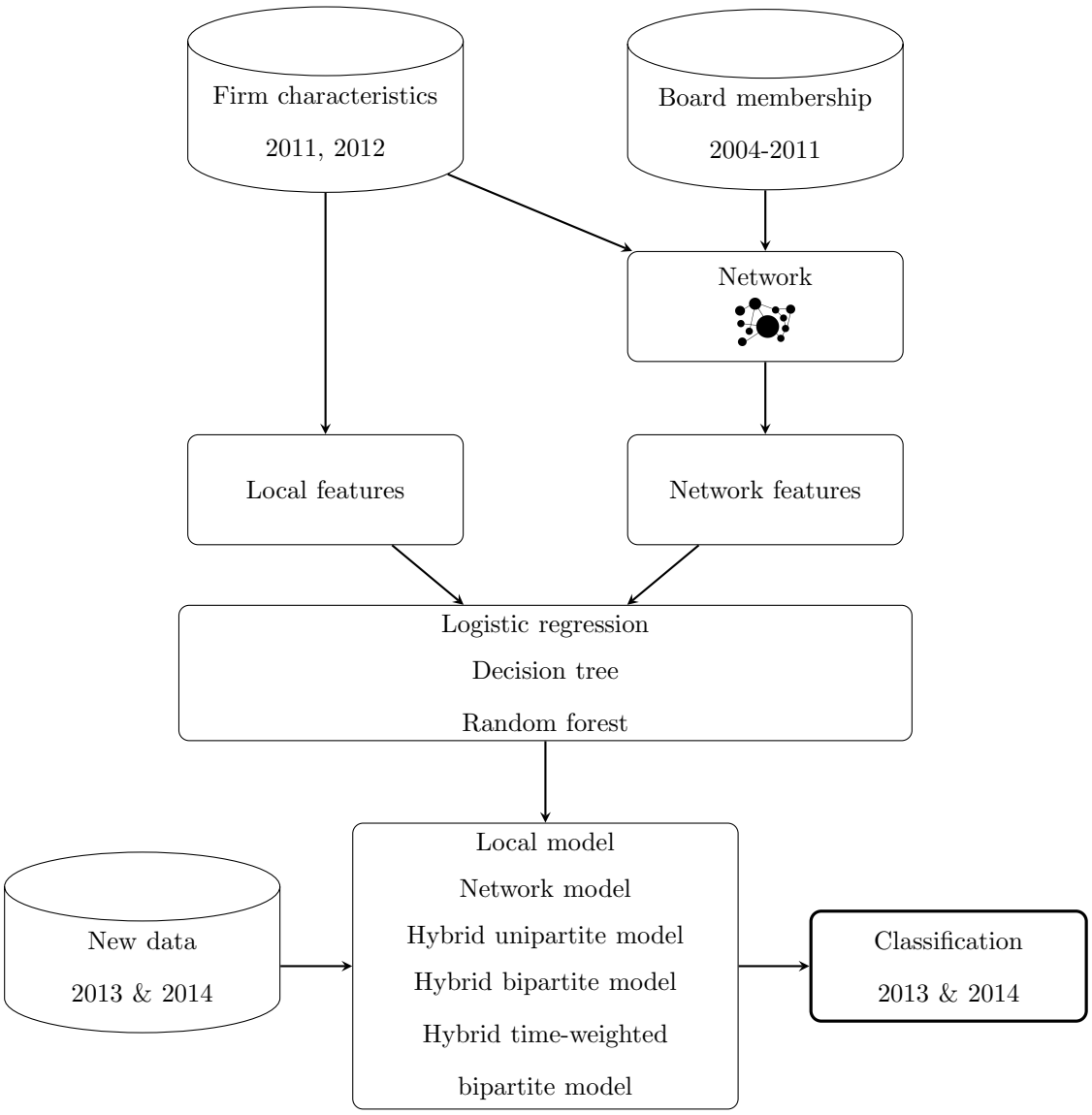
the unipartite network and is referred to as the network model. Thirdly, we construct a hybrid model using both local variables and network variables extracted from the unipartite network. Similarly, models four and five combine local variables with network variables from the unweighted and time-weighted bipartite network respectively. The whole methodology is visualized in Figure 3.

4. Results and discussion

4.1. Is there support for homophily?

Homophily in social networks occurs when the likelihood that two similar people are connected is larger than the likelihood that two random people are connected (McPherson et al., 2001). Our study finds support for homophily indicating that firms who have a low effective tax rate are more connected to each other by means of shared board members than randomly expected (p-value < 0.0001 using a proportions test with continuity correction). As a proxy for homophily, we can also study whether the network is dyadic and heterophobic (Park & Barabsi, 2007; Baesens et al., 2015). The network is dyadic if low-tax firms are more densely connected than randomly, and it is heterophobic if low-tax firms are less connected to other firms than expected

Figure 3: Methodology



if they were randomly connected. With a dyadicity of $D = 0.88$ (dyadicity is supported if $D > 1$) and a heterophilicity of $H = 0.77$ (heterophilicity is supported if $H < 1$), we can conclude that there is only support for heterophilicity. These findings encourage further analysis by means of social network analytics for low-tax prediction.

4.2. 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 leading to a selected subset of the variables. This feature selection process was based on the Akaike information criterion (AIC) measure and applied both in a stepwise forward and backward manner. Afterwards, remaining non-significant variables (p-value > 0.10) were consecutively omitted. As can be observed from Table 2, the hybrid unweighted bipartite model performs best in terms of AUC.

Table 2: Performance of the logistic regression models in terms of accuracy and AUC

	2013		**2014**	
	Accuracy	**AUC**	**Accuracy**	**AUC**
Local model	88.89%	66.27%	89.61%	68.82%
Network unipartite model	88.97%	60.15%	89.69%	55.19%
Hybrid unipartite model	88.73%	67.10%	89.61%	68.17%
Hybrid unweighted bipartite model	90.09%	83.99%	90.01%	82.87%
Hybrid time-weighted bipartite model	89.53%	83.94%	89.29%	83.32%

We furthermore note that it significantly outperforms the local, network and hybrid unipartite model (p-values < 0.0001 using the test of De-

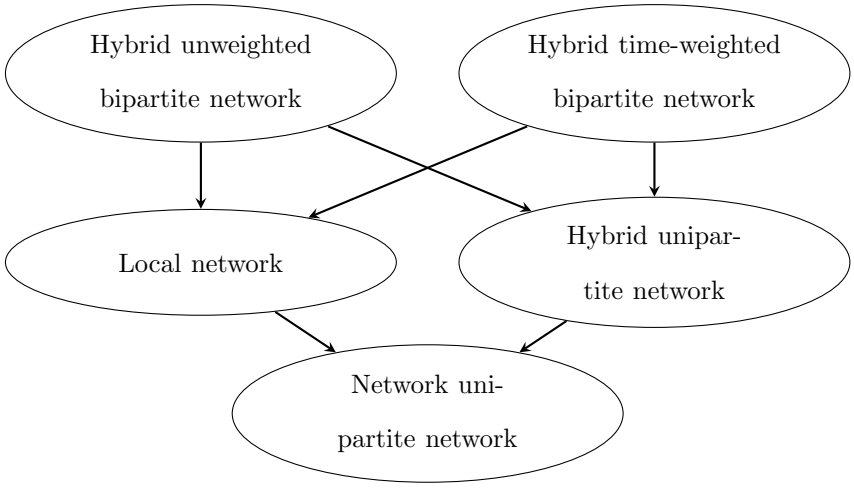


Figure 4: Domination graph (Rossetti et al., 2016) of random forest models based on pairwise comparison of AUC values (DeLong et al., 1988). Arrows indicate a significant performance improvement in AUC at a 0.1% significance level.

Long et al. (1988)). 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 (p-values > 0.45 using the test of DeLong et al. (1988)). 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. The pruning parameter is tuned by means of a ten-fold cross-validation repeated three times on the training set. Again, the hybrid unweighted bipartite model performs the best (AUC = 0.7636 for 2013 and AUC = 0.6281 for 2014) thereby significantly outperforming the local model (AUC = 0.5998 for 2013 and AUC = 0.7465 for 2014). However, the network unipartite and hybrid

unipartite model are performing badly with AUCs equal to 0.5. Nevertheless, we observe a benefit in modelling non-linear effects. Therefore, we train random forests next. 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. Table 3 shows that the hybrid bipartite models clearly outperform the other models in terms of AUC. The local and hybrid unipartite models perform slightly worse but still surpass the network unipartite model. Figure 4 illustrates how the models compare to each other in terms of significant improvement in AUC. Furthermore, all models show an improvement towards their logistic regression counterpart. Additionally this comparison is illustrated by means of ROC curves in Figures 5a and 5b for 2013 and 2014 respectively.

Table 3: Performance of the random forest models in terms of accuracy and AUC

	2013		**2014**	
	Accuracy	**AUC**	**Accuracy**	**AUC**
Local model	89.37%	76.83%	90.09%	74.89%
Network unipartite model	88.89%	60.18%	89.61%	56.11%
Hybrid unipartite model	89.05%	74.96%	89.77%	74.74%
Hybrid unweighted bipartite model	89.77%	84.31%	90.33%	83.06%
Hybrid time-weighted bipartite model	89.13%	84.12%	89.69%	83.33%

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 4 summarizes both metrics at a cut-off of 50%

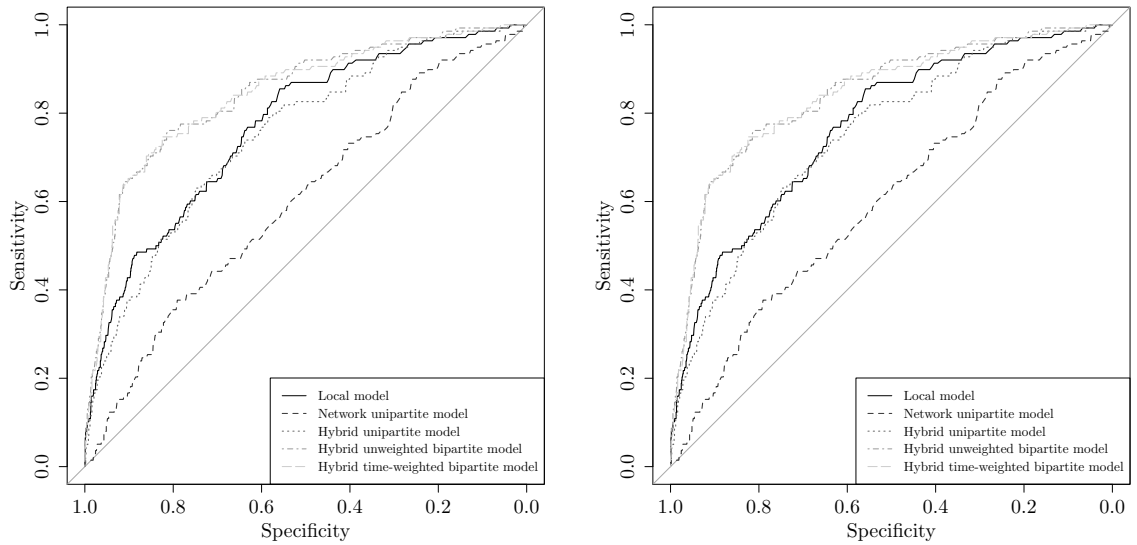
and an adapted cut-off so that the ratio of low-tax firms in the predictions equals the ratio of low-tax firms in the validation sets. As such, the adapted cut-off will classify the 11% and 10%, for 2013 and 2014 respectively, most likely to be low-tax firms as low-tax in fact. This metric will inform 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.

Table 4: 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 validation sets.

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

4.3. 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



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

Figure 5: 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.

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.

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 6. We observe that there are three important characteristics of a firm: a lower EBITDA, a non-missing value for

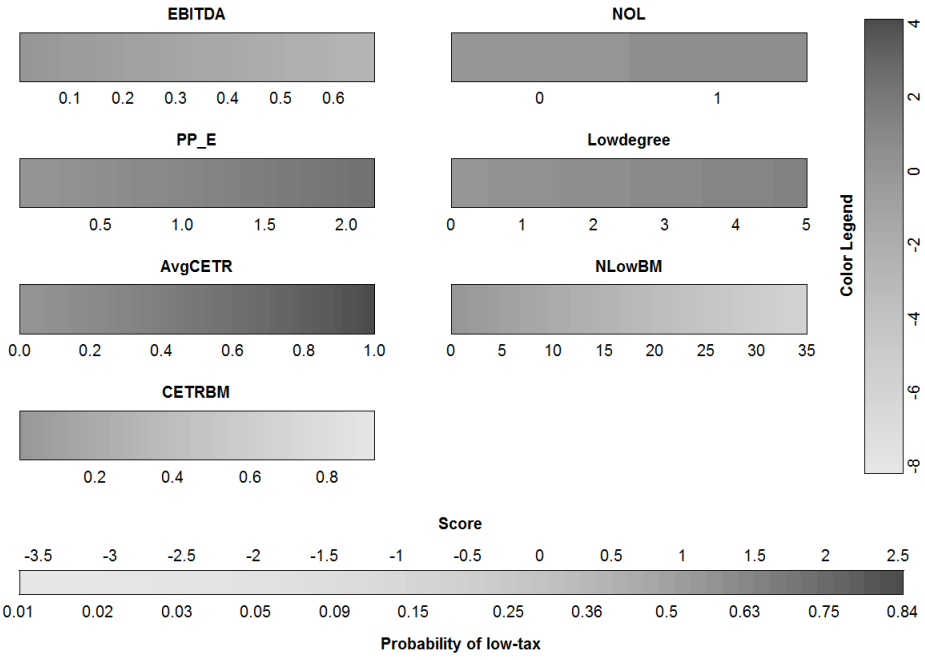


Figure 6: Colored nomogram. The color indicates the extent to which a variable contributes to the probability of being a low-tax firm, and can be converted to points by means of the Color Legend (on the right). To calculate the final probability, all points can be summed and converted by means of the Score bar (at the bottom). This visualization was created based on the work of Van Belle & Van Calster (2015).

its tax loss carry-forward and a higher PP&E lead to an increased probability of being a low-tax firm. For the network characteristics, a higher number of neighboring low-tax firms, a higher average CETR of a firm's neighbors, a lower number of board members who are not connected to low-tax firms, and a lower average CETR of the neighbors of a firm's board members, 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 captured by the logistic regression model. When we, in addition, take a closer at the hybrid unipartite model, we observe a positive effect of betweenness.

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 (Neuman, 2014). This increases support for the idea of a valuable information flow on tax strategies between firms through 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. Figure 7 illustrates the mean decrease in node impurity when we split a tree based on a certain variable. 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. 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.

4.4. 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 reduces its potential given that good results were previously obtained in the fraud detection domain (Van Vlasselaer et al., 2017). 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

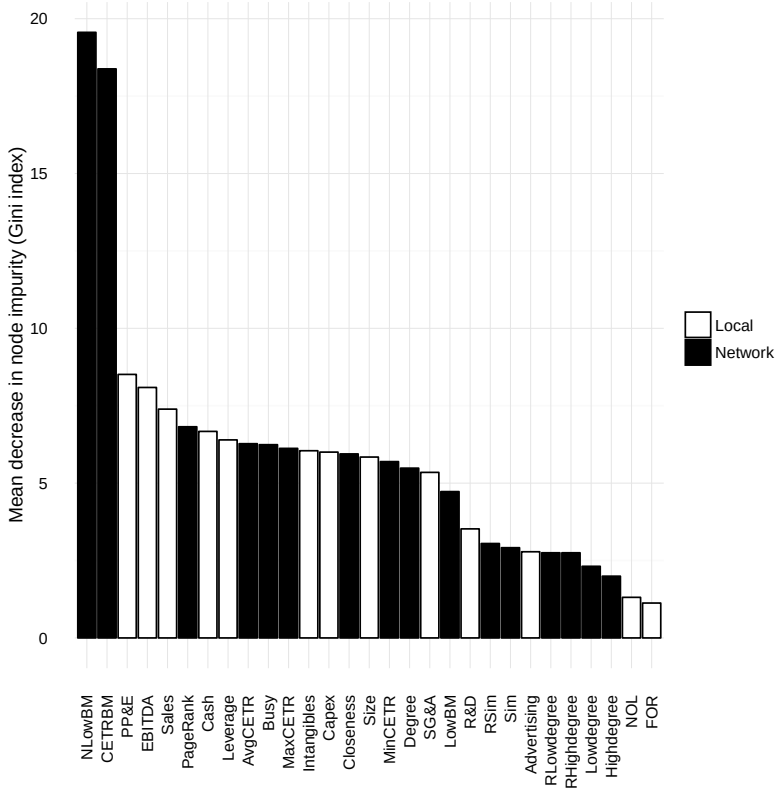


Figure 7: Mean decrease in node impurity measured by the Gini index if a particular variable is removed from the variable set.

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 perform better, e.g. artificial neural networks, support vector machines, etc.

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 researched. Our hybrid bipartite random forest model performed best with an 7% 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. Additionally, we gained insights that can assist companies in 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, 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 also benefit from the ability to better identify tax evading firms.

Acknowledgements

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

References

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. doi:10.1016/j.jacceco.2015.02.003.

Armstrong, C. S., Blouin, J. L., & Larcker, D. F. (2012). The incentives for tax planning. *Journal of Accounting and Economics*, *53*, 391–411. doi:10.1016/j.jacceco.2011.04.001.

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.

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).

Bizjak, J., Lemmon, M., & Whitby, R. (2009). Option backdating and board interlocks. *Review of Financial Studies*, *22*, 4821–4847. doi:10.1093/rfs/hhn120.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. doi:10.1023/A:1010933404324.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Chapman and Hall/CRC.

Brown, J. L., & Drake, K. D. (2014). Network ties among low-tax firms. *The Accounting Review*, *89*, 483–510. doi:10.2308/accr-50648.

Bruynseels, L., & Cardinaels, E. (2014). The audit committee: Management watchdog or personal friend of the CEO? *The Accounting Review*, *89*, 113–145. doi:10.2308/accr-50601.

Cashman, G. D., Gillan, S. L., & Jun, C. (2012). Going overboard? On busy directors and firm value. *Journal of Banking & Finance*, *36*, 3248–3259. doi:10.1016/j.jbankfin.2012.07.003.

Chiu, P.-C., Teoh, S. H., & Tian, F. (2013). Board interlocks and earnings management contagion. *The Accounting Review*, *88*, 915–944. doi:10.2308/accr-50369.

Christensen, D. M., Dhaliwal, D. S., Boivie, S., & Graffin, S. D. (2015). Top management conservatism and corporate risk strategies: Evidence from managers' personal political orientation and corporate tax avoidance. *Strategic Management Journal*, *36*, 1918–

1938. doi:10.1002/smj.2313.

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).

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. doi:10.2307/2531595.

Desai, M. A., & Dharmapala, D. (2006). Corporate tax avoidance and high-powered incentives. *Journal of Financial Economics*, *79*, 145–179. doi:10.1016/j.jfineco.2005.02.002.

Dyreng, S. D., Hanlon, M., & Maydew, E. L. (2010). The effects of executives on corporate tax avoidance. *The Accounting Review*, *85*, 1163–1189. doi:10.2308/accr.2010.85.4.1163.

Gallemore, J., & Labro, E. (2015). The importance of the internal information environment for tax avoidance. *Journal of Accounting and Economics*, *60*, 149–167. doi:10.1016/j.jacceco.2014.09.005.

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. doi:10.2308/accr-50678.

Gupta, S., & Newberry, K. (1997). Determinants of the variability in corporate effective tax rates: Evidence from longitudinal data. *Journal of Accounting and Public Policy*, *16*, 1–34. doi:10.1016/S0278-4254(96)00055-5.

Hanlon, M., & Heitzman, S. (2010). A review of tax research. *Journal of Accounting and Economics*, *50*, 127–178. doi:10.1016/j.jacceco.2010.09.002.

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. doi:10.1007/978-1-4419-8462-3_1.

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. doi:10.1111/j.1468-5957.2011.02276.x.

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. doi:10.1007/978-3-662-04599-2_11.

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. doi:10.1016/j.jaccpubpol.2010.09.003.

Larcker, D. F., So, E. C., & Wang, C. C. (2013). Boardroom centrality and firm performance. *Journal of Accounting and Economics*, *55*, 225–250. doi:10.1016/j.jacceco.2013.01.006.

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. doi:10.1016/j.ejor.2015.05.030.

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. http://dl.acm.org/citation.cfm?id=1314532.

Martin, S., Brown, W. M., Klavans, R., & Boyack, K. W. (2011). OpenOrd: an open-source toolbox for large graph layout. In *Visualization and Data Analysis 2011, San Francisco Airport, CA, USA, January 24-25, 2011.* SPIE Proceedings (p. 786806). doi:10.1117/12.871402.

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. doi:10.2308/accr-10215.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415–444. http://www.jstor.org/stable/2678628.

Minnick, K., & Noga, T. (2010). Do corporate governance characteristics influence tax management? *Journal of Corporate Finance*, *16*, 703–718. doi:10.1016/j.jcorpfin.2010.08.005.

Neuman, S. S. (2014). Effective tax strategies: It's not just minimization. doi:10.2139/ssrn.2496994. Available at SSRN: https://ssrn.com/abstract=2496994 (Last revised: 09/17/2014).

Omer, T. C., Shelley, M. K., & Tice, F. M. (2014). Do well-connected directors affect firm value? *Journal of Applied Finance*, *24*, 17–32. https://ssrn.com/abstract=2665654.

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. Available at SSRN: https://ssrn.com/abstract=2379151 (Last revised: 02/09/2016).

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

Park, J., & Barabsi, A.-L. (2007). Distribution of node characteristics in complex networks. *Proceedings of the National Academy of Sciences*, *104*, 17916–17920. doi:10.1073/pnas.0705081104.

Rego, S. O. (2003). Tax-avoidance activities of u.s. multinational corporations. *Contemporary Accounting Research*, *20*, 805–833. doi:10.1506/VANN-B7UB-GMFA-9E6W.

Robinson, J. R., Xue, Y., & Zhang, M. H. (2012). Tax planning and financial expertise in the audit committee. doi:10.2139/ssrn.2146003. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2146003 (Last revised: 08/31/2012).

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). doi:10.1145/2959100.2959176.

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

Slemrod, J. (2004). *The Economics of Corporate Tax Selfishness*. Working Paper 10858 National Bureau of Economic Research. doi:10.3386/w10858.

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).

Therneau, T., Atkinson, B., & Ripley, B. (2015). *rpart: recursive partitioning and regression trees*. R package version 4.1-10. https://CRAN.R-project.org/package=rpart.

Van Belle, V., & Van Calster, B. (2015). Visualizing risk prediction models. *PLoS ONE*, *10*, e0132614. doi:10.1371/journal.pone.0132614.

Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2017). GOTCHA! Network-based fraud detection for social security fraud. *Management Science*, *forthcoming*. doi:10.1287/mnsc.2016.2489.

Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. *Applied Soft Computing*, *14*, 431–446. doi:10.1016/j.asoc.2013.09.017.

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 1, 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[1]				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****

[1]LowBM was excluded after feature selection presumably because 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.

WBusy					Not included

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

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)