DEPARTMENT OF ENGINEERING MANAGEMENT

Data Mining for Fraud Detection using Invoicing Data: A Case Study in Fiscal Residence Fraud

David Martens, Enric Junqué de Fortuny & Marija Stankova

UNIVERSITY OF ANTWERP
Faculty of Applied Economics

City Campus
Prinsstraat 13, B.226
B-2000 Antwerp
Tel. +32 (0)3 265 40 32
Fax +32 (0)3 265 47 99
www.uantwerpen.be

FACULTY OF APPLIED ECONOMICS

DEPARTMENT OF ENGINEERING MANAGEMENT

Data Mining for Fraud Detection using Invoicing Data: A Case Study in Fiscal Residence Fraud

David Martens, Enric Junqué de Fortuny & Marija Stankova

RESEARCH PAPER 2013-026
OCTOBER 2013

University of Antwerp, City Campus, Prinsstraat 13, B-2000 Antwerp, Belgium
Research Administration – room B.226
phone: (32) 3 265 40 32
fax: (32) 3 265 47 99
e-mail: joeri.nys@uantwerpen.be

The research papers from the Faculty of Applied Economics are also available at www.repec.org (Research Papers in Economics - RePEc)

D/2013/1169/026

Data Mining for Fraud Detection using Invoicing Data

A Case Study in Fiscal Residence Fraud

David Martens · Enric Junqué de Fortuny · Marija Stankova

Abstract This paper describes a methodology to efficiently build predictive fraud detection models based on payment transaction data. More specifically, a network learning technique is applied using invoicing data from and to foreign companies. A network is created among foreign companies, where two companies are connected if they have sent an invoice to (or received an invoice from) the same Belgian company. These connections are weighted, taking into account the number of shared Belgian companies and the popularity of the Belgian company that links the foreign companies. Data mining techniques are applied to predict residence fraud committed by foreign companies. Our empirical results show that the obtained models are indeed able to discriminate between fraudulent and non-fraudulent companies, with an AUC up to 79%. The superiority of our proposed method is shown by comparing its results to a support vector machine trained on the same transacitonal data (including SVD and balancing of the dataset).

1 Introduction

Fraud is a large scale problem that affects a multitude of entities: the public sector, the private sector, individuals and even charity organizations [6]. The overall impact and scale of fraud is very difficult to measure since most of it remains undetected and only estimates can be made. Governments are a frequent target of fraudsters that undermine the system and abuse its benefits, grants and tax programs. The abuse of the tax system is the most costly fraud type [6], with estimates of losses going up to few billions for the governments of Australia, US or UK and even trillion euros in the European Union (see Table 1). These losses have a direct financial impact on the individuals as well. For example in the UK, fraud against the public sector is estimated

Applied Data Mining Research Group, Faculty of Applied Economics, University of Antwerp, Belgium
E-mail: David.Martens@uantwerp.be

to be £20.6 billion per year and directly costs every adult about £1,460 annually[1]. The money lost in fraud could mean more budget cuts for the government, tax increases, less investments in the public sector (such as new roads, hospitals, schools, etc.) and eventually a slower economy altogether. In Table 1 we can see the immediate financial losses for several industries in different regions. The numbers are striking and in the range of a few billions per industry.

In this paper, we consider company residence fraud, which is one of the many fiscal and social frauds that a government is faced with. It basically entails the problem that a company is fraudulently listed in another company for fiscal or social advantages. In the next section we describe the data obtained, being a set of known frauds and a large dataset of invoicing data from Belgian companies to foreign companies and vice verca. Section 3 describes the methodology to go from this 'big data' to predictive models that can find fraudulent foreign companies. The results are described in Section 4, whereas Section 5 concludes the paper.

Table 1 Size of different frauds per region.

Type	Region	Amount	Year	Source
Credit Card	USA	US$ 3.5 billion	2012	CyberSource (subsidiary of Visa)
	EU	€ 1.16 billion	2011	European Central Bank
	UK	£ 388 million	2012	National Fraud Authority
	Australia	AU$ 261 million	2012	Australian Payments Clearing Ass.
	Canada	CA$ 439 million	2012	Canadaian Bankers Ass.
Tele-communications	Worldwide	US$ 46.3 billion	2013	Comm. Fraud Control Ass. (CFCA)
	UK	£ 953 million	2011	National Fraud Authority
Insurance	UK	£ 2.1 billion	annually	National Fraud Authority
	USA	US$ 80 billion	annually	Coalition Against Insurance Fraud
	EU	€ 110 billion	annually	Insurance Europe
	Australia	AU$ 1.4 billion	annually	Australian Institute of Criminology
	Canada	CA$ 1 billion	annually	Insurance Bureau of Canada
Tax Fraud	UK	£ 14 billion	2010-11	National Fraud Authority
	EU	€ 1 trillion	annually	European Commission
	USA	US$ 100 billion	annually	Congressional Research Service

2 Data

Anonymized payment data is obtained from a total of 2,745,478 Belgian companies and 873,702 foreign companies. Two types of payments are considered: (1) incoming invoices, which are transactions from foreign to Belgian companies, and (2) outgoing invoices, from Belgian to foreign companies. Three different datasets were constructed

[1] http://archive.audit-commission.gov.uk/auditcommission/sitecollectiondocuments/Downloads/20121107-ppp2012.pdf

from these payment data: a dataset of incoming invoices, a dataset of outgoing invoices and a third dataset where we merged both the incoming and outgoing invoices. The target variable to predict is whether a foreign company is committing residency fraud or not. A problem that arises is the skewness of the data, seen that out of the 873,702 foreign companies that transact with Belgian companies, only 62 are positive cases.

Additional statistics for the datasets are shown in Table 2. Note that there can be multiple transactions between two companies. Hence in Table 2, both the total number of transactions and the number of unique transactions between Belgian and foreign companies are given. The latter counts only the transactions where the invoice sender and recipient are unique. Figure 1 shows the degree distributions (number of invoices) that a foreign company sends and receives. We can see that most foreign companies that transact with Belgian companies, typically only send or receive invoices to few Belgian companies. For example, a Belgian company that transacts with foreign companies, will receive on average 9.84 invoices. Only a few foreign companies send or receive invoices to thousands of Belgian companies, indicated by the right tail of the distributions.

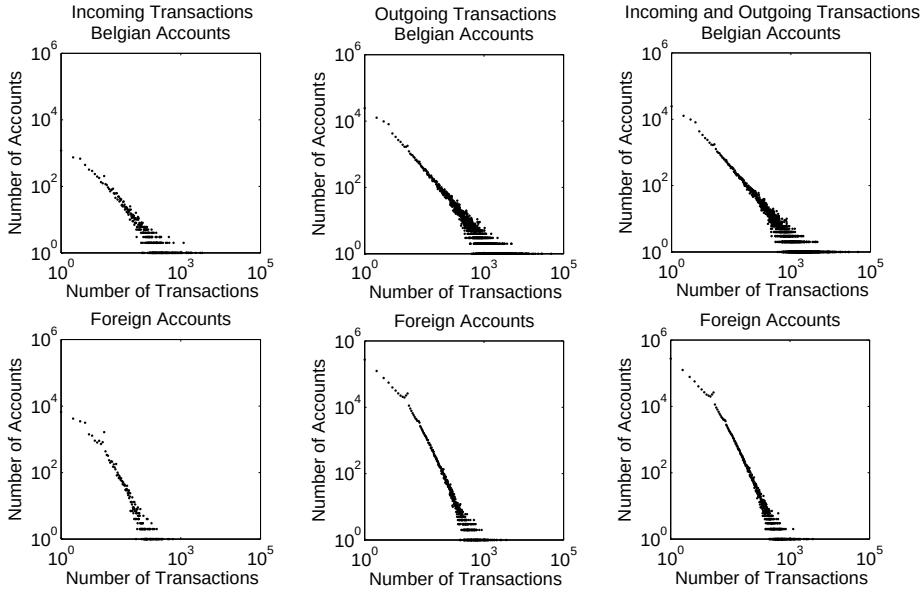

Fig. 1 Number of transactions per account for the payment datasets.

3 Methodology

In this work, we applied the methodology proposed by Martens and Provost [5, 4] where payment data from a large bank are used to create a network among 1 million consumers, based on their payments to a total of 6 million merchants. In their work, they show the suitability of this method to predict which consumers are likely interested in a financial product.

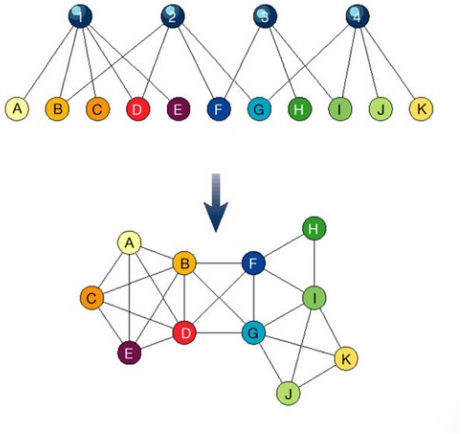

Fig. 2 From invoicing data between foreign (A-K) and Belgian (1-4) companies to a network among foreign companies that are connected if they receive (send) an invoice to the same Belgian company.

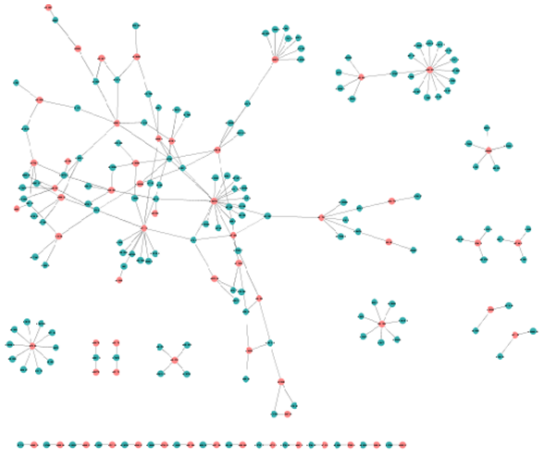

Fig. 3 Structure of the payment network based on the incoming transactions.

Table 2 Statistics for the three payment datasets.

	Incoming	**Outgoing**	**Combined**
Number of transactions	251,198	6,551,512	6,802,710
Number of unique transactions	73,753	1,955,912	2,029,641
Number of Belgian accounts	7,495	107,345	108,753
Number of Foreign accounts	30,541	858,131	858,703
Average number of transactions per Belgian account	9.84	18.22	18.66
Average number of transactions per foreign account	2.41	2.28	2.36

Fig. 4 Structure of the payment network based on the outgoing transactions.

More specifically, the proposed targeting method constructs a 'pseudo-social network' (PSN) where two companies are linked if they receive an invoice from the same entity (the same applies for sending invoices). This idea is illustrated in Figure 2. Figures 3-4 show the network structure of the fraudulent foreign companies (red nodes), connected to other foreign companies with which they share at least one Belgian company. It nicely shows that fraudulent companies tend to be connected to each other and the same other foreign companies, already showing evidence that fraudulent companies indeed tend to send invoices to the same Belgian companies. Such cliquing behavior is exactly what relational learners pick up on.

The connections are weighted according to the number of shared Belgian companies: if foreign companies F_1 and F_2 both receive invoices from the Belgian companies B_1 and B_2, their connection will be twice as strong as the connection between two foreign companies that share only one Belgian company. Additionally, the popularity of a Belgian company is considered in the weighting scheme: two foreign companies that both receive an invoice from a Belgian company where no one else receives an invoice from, are likely very similar. On the other hand, the fact that two foreign companies share a Belgian company where thousands of other foreign companies also receive an invoice from (e.g. Coca-Cola) is much less telling for their similarity, and hence should result in a low weight connection. For more details on the possible weighting schemes, we refer to [8]. The tangens hyperbolicum and tunable beta function are used to downweight 'popular' Belgian companies.

On the resulting network, relational learners such as weighted voting relational learner (wvRN) and network-only link based (nLB) classifiers can be applied that can predict if a foreign company is fraudulent [3]. The basic assumption of the relational learner is that if a company is linked to a fraudulent company, it is likely also fraudulent, and the stronger the connection with the fraudulent company, the more likely it is also committing fraud. The combination of certain design choices can be implemented in a very fast routine that result in a linear model, as described and proven in [8].

4 Empirical Study

4.1 Performance measurement

In the experimental setup, 80% of the data is used as training data to build the predictive model, which is subsequently tested on the remaining 20% of the data. As such, we obtain an estimate of how well the model would perform on new, unseen data instances. To ensure we don't get specifically good or bad results by chance, this procedure is repeated five times with a different random splitup in training and test set. We report the average result over these five experiments (5-fold cross testing). The models are evaluated using the area under the ROC curve, as accuracy is known to be a misguided metric [7]: a model predicting everyone to be non-fraudulent will have an accuracy of 99.9999%, even though the model is clearly useless. A random model yields an AUC of 50%, a perfect model has an AUC of 100%. Additionally, we also report the recall and precision. Recall is the percentage of the known frauds that are also discovered by the predictive model, whereas precision is the percentage of the predicted frauds that are also known frauds.

4.2 Benchmark techniques

The proposed technique is compared to several state-of-the-art classification techniques. The first is the popular support vector machine (SVM) with linear kernel. The data is represented in matrix format, where each foreign company corresponds to a row, and each possible Belgian company corresponds to a column. If foreign company F_i receives (sends) an invoice from Belgian company B_j, the value of element x_{ij} is set to 1 (we only consider binary values). The SVM can be applied to such high-dimensional data (with 100 thousands of variables) through a concept known as regularization, which basically penalizes for complexity. To reduce the dimensionality, we perform a singular value decomposition and apply a linear SVM on the reduced matrix. To deal with the very skewed dataset (only 62 fraud cases), we also consider oversampling these positive cases (repeating them several times). Finally, we also apply Vowpal Wabbit [2], a technique developed at Microsoft and Yahoo! Research Labs designed to be able to deal with huge datasets, and a naive Bayes implementation tailored for big data, named Big Bayes [1].

4.3 Results

The results are summarized in Tables 3 and 4. Figure 5 visualizes the AUC of the resulting model, with the standard deviation of the results over the five test folds. The tables and figure show that the proposed PSN method performs very well, with an AUC up to 79,6% (using 5-fold cross-testing). The best results are obtained when we use the daatset of outgoing invoices, from Belgian companies to foreign companies. The best performing technique is obtained from applying the nLB relational learner on the pseudo-social network, closely followed by the wvRN classifier. An additional advantage of the wvRN relational learner is that it provides a linear model, where each Belgian company receives a coefficient. As such, we can also investigate the Belgian

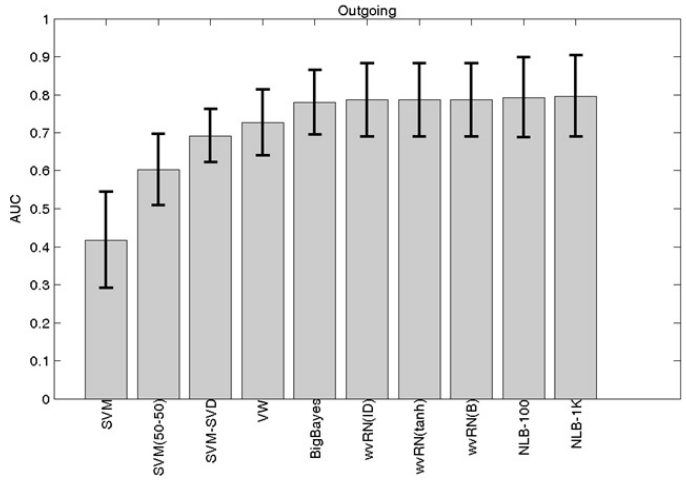

Fig. 5 Results in terms of AUC for the dataset of the outgoing invoices (from Belgian to foreign companies).

companies with a high coefficient as this is an indication that they transact frequently with fraudulant foreign companies.

Table 3 Results in terms of accuracy, recall and precision.

Number of most fraudulant cases to consider	Accuracy	Recall	Precision
100	99,90%	10,40%	1,20%
500	99,70%	27,50%	0,60%
1 000	99,40%	32,80%	0,36%
10 000	94,20%	62,20%	0,07%

5 Conclusion

Invoicing data holds very useful fine-grained information about foreign companies. Such data is often summarized in a few variables, such as the number of Belgian companies it transacts with, mainly because of the inability of existing classification techniques to deal with hundreds of thousands of variables. We show that the PSN method is able to leverage the fine-grained data in a very scalable manner (the computational time to process this data is in the order of minutes), and holds much promise for building predictive fraud detection models.

Table 4 Results in terms of AUC.

	Incoming	**Outgoing**	**Combined**
NLB-1K	75,60%	**79,60%**	78,00%
NLB-100	74,80%	79,20%	78,10%
wvRN(B)	75,80%	78,60%	77,50%
wvRN(tanh)	75,90%	78,60%	77,70%
wvRN(ID)	75,80%	78,60%	77,70%
BigBayes	75,80%	77,90%	76,90%
VW	75,40%	72,70%	74,20%
SVM-SVD	63,70%	69,10%	72,10%
SVM(50-50)	63,70%	60,30%	59,80%
SVM	53,80%	41,80%	45,00%

6 Acknowledgments

We are very grateful to the Belgian government for allowing us to work on the data and the useful feedback. We also would like to thank SAS for the fruitful discussions and feedback. Note that these models or techniques are not necessarily used in production.

References

1. E. Junqué de Fortuny, F. Provost, and D. Martens. Is bigger data really better? *Big Data Journal*, 2013.
2. J. Langford, L. Li, and A. Strehl. Vowpal Wabbit, 2007.
3. S. A. Macskassy and F. Provost. A simple relational classifier. 2003.
4. D. Martens and F. Provost. Methods, computer-accessible medium and systems for construction of and inference with networked data, for example, in a financial setting. Patent Application WO2011112981 A3, March 2011.
5. D. Martens, F. Provost, J. Clark, and E. Junqué de Fortuny. Pseudo-social network targeting from consumer transaction data. Working paper CeDER-11-05, New York University - Stern School of Business, 2011.
6. National Fraud Authority. Annual fraud indicator 2013. 2013.
7. F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing classifiers. In *15th International Conference on Machine Learning*, 1998.
8. M. Stankova, D. Martens, and F. Provost. Classification over bipartite graphs through projection. 2013.