TARC
Tax Administration Research Centre

Detecting anomalies in heterogeneous population-scale VAT networks

ANGELOS ALEXOPOULOS, UNIVERSITY OF CAMBRIDGE

PETROS DELLAPORTAS, UNIVERSITY COLLEGE LONDON

STANLEY GYOSHEV, UNIVERSITY OF EXETER

CHRISTOS KOTSOGIANNIS, TARC, UNIVERSITY OF EXETER

SOFIA C OLHEDE, ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

TRIFON PAVKOV, NATIONAL REVENUE AGENCY (BULGARIA)

DETECTING ANOMALIES IN HETEROGENEOUS POPULATION–SCALE VAT NETWORKS

Angelos Alexopoulos[a,1], Petros Dellaportas[b,c,d], Stanley Gyoshev[ℓ], Christos Kotsogiannis[e,f,k], Sofia C. Olhede[b,g], Trifon Pavkov[ℓ,h]

Abstract: Anomaly detection in network science is the method to determine aberrant edges, nodes, subgraphs or other network events. Heterogeneous networks typically contain information going beyond the observed network itself. Value Added Tax (VAT, a tax on goods and services) networks, defined from pairwise interactions of VAT registered taxpayers, are analysed at a population–scale requiring scalable algorithms. By adopting a quantitative understanding of the nature of VAT–anomalies, we define a method that identifies them utilising information from micro–scale, meso–scale and global–scale patterns that can be interpreted, and efficiently implemented, as population–scale network analysis. The proposed method is automatable, and implementable in real time, enabling revenue authorities to prevent large losses of tax revenues through performing early identification of fraud within the VAT system.

[a]MRC Biostatistics Unit, University of Cambridge, University Forvie Site, Robinson Way, Cambridge CB2 0SR, UK. Email: angelos@mrc-bsu.cam.ac.uk.
[b]Department of Statistical Science, University College, Gower Street, London, WC1E 6BT, UK. Email: p.dellaportas@ucl.ac.uk.
[c]Department of Statistics, Athens University of Economics and Business, Greece.
[d]The Alan Turing Institute, London, UK.
[e]Department of Economics, University of Exeter Business School, Streatham Court, Rennes Drive, Exeter EX4 4PU, England, UK. Email: c.kotsogiannis@exeter.ac.uk.
[f]ESIfo, Munich, Germany.
[g]Institute of Mathematics, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland. Email: sofia.olhede@epfl.ch.
[h]National Revenue Agency, Sofia, Bulgaria. Email: t.pavkov@nra.bg.
[k]Tax Administration Research Centre (TARC), University of Exeter.
[ℓ]Department of Finance, University of Exeter Business School, Streatham Court, Rennes Drive, EX4 4PU, England, UK. Email: s.gyoshev@exeter.ac.uk.

Significance Statement: The Value Added Tax (VAT) suffers from fraud problems arising through fictitious network transactions. We develop a scalable and automatable anomaly detection method that combines the VAT network structure with machine learning and applies it to VAT data from the Bulgarian National Revenue Agency (BNRA). BNRA applies risk-based rules on over 300,000 taxpayers to identify VAT fraud. The prioritisation process initially selects 15,000 which is narrowed down, with additional sample refinement, to 500 and, through auditing, approximately 100 taxpayers are considered as likely members of VAT fraud. The proposed method identifies, through a one stage process, 100 VAT fraudsters by considering only 200 taxpayers for auditing. The method generates savings on scarce resources and improves tax compliance and revenue yield.

1 Introduction

Value Added Tax (VAT) is a major source of revenue for, remarkably,[1] over 160 countries. VAT is a consumption tax in the sense that the VAT collected through the supply chain is the VAT paid by the consumers in the place where the good is consumed. Underlying VAT therefore there is an "invoice-credit" mechanism where the net tax liability of a business[2] is calculated by subtracting from the sales the aggregate value of VAT paid on invoices for the inputs used in production. The "invoice–credit" mechanism requires sellers along the production chain to provide invoices to their buyers showing the amount of VAT that was paid on a given transaction. The fractional revenue collection on the value added that is generated at every stage of the production chain is remitted to the appropriate revenue authority. The business-to-business (B2B) transactions and the VAT "invoice–credit" mechanism de-facto create a *network* through which businesses are interacting within and across economic sectors.

Despite its remarkable rise as a tax innovation, it is universally recognised that the current VAT system has both weaknesses and vulnerabilities (Keen and Smith, 2006) making it not fit for purpose in a modern technology-based economy (Ebrill et al., 2001). Like any tax, VAT is vulnerable to fraud, but the "invoice–credit" mechanism offers unique opportunities for abuse, an issue that has become a major concern in the European Union and its Member States (Keen and Smith, 2006). *Missing Trader Intra-Community (MTIC) fraud* (famously also known as *carousel fraud*) is the most challenging type of VAT fraud and is a consequence of zero-rating under VAT: The fraudsters, exploiting the fact that exports are zero-rated, claim a refund of VAT paid on purchases even though output VAT has not yet been collected in the country the goods have been exported to. This structural element of VAT has been eloquently described as VAT's Achilles heel (Keen and Smith, 2006).[3] VAT fraud is not unique to cross-border B2B transactions but arises within the domestic VAT network too. The analysis and the method developed in this paper captures all elements of VAT network fraud.

Fraud in the VAT system corresponds to anomalous behaviour in the web of interactions between traders within the network of traders. This, in parts, is associated with the anomalous behaviour of individual traders (or nodes) in the network of interactions. However, as VAT fraud requires the interaction of multiple B2B traders, it corresponds to a communal behaviour of interactions in a group of nodes. To isolate potential fraudulent behaviour requires modeling of both individual behaviour of nodes and communal

---

[1]For the remarkable rise of VAT, and what has shaped its adoption, see, for example, Keen and Lockwood (2010).

[2]Throughout trader and business are used interchangeably. A requirement for a trader/business to claim VAT is that they are registered for VAT with the revenue authority.

[3]MTIC fraud is not unique to the European Union but it is also of relevance to countries where fiscal checks at the physical borders have been relaxed following trade agreements.

behaviour of groups using network science. This also requires the satisfaction of the constraint that if the method is to be applied to a significant portion of data then the method must be both robust and computationally efficient.

VAT fraud presents a significant threat to society's welfare as it erodes tax revenues but also impedes the "level playing field" putting legitimate businesses into a disadvantaged position. Identifying and combating VAT fraud before it occurs is therefore important, particularly so when it involves missing traders: Once the fraud materialises and any VAT claim to the revenue authority has been refunded, tracking down the fraudsters and recovering the lost revenues is, in most cases, an impossibility. Given the high volume of transactions across sectors and between businesses, risk–based profiling of VAT claims is required coupled with early, timely, and preferably automatic, fraud-detection.

The objective of this paper is to develop an anomaly detection and automatable method suitable for dealing with a population–scale and heterogeneous network such as the one constructed from VAT transactions. The possibility of automating the detection of VAT–fraud is part of a larger current international research theme seeking to use large scale data sets to improve tax (and social) policy (Lazer et al., 2009; Athey, 2017; Lazer et al., 2020) and our understanding of human interactions (Jackson and Wolinsky, 1996; Margetts and Dorobantu, 2019).

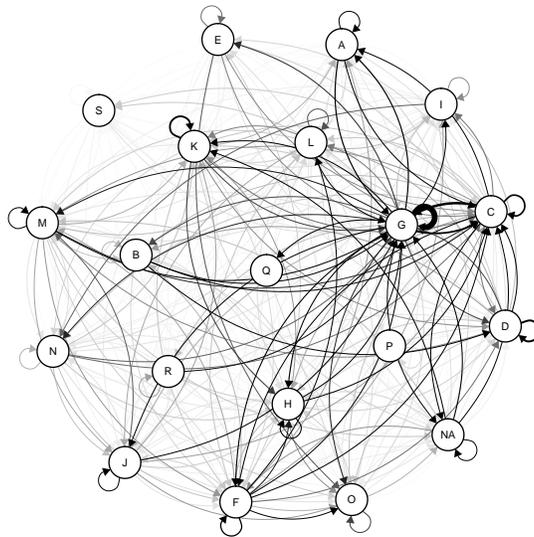Figure 1: The directed weighted network of sector specific VAT transactions that were conducted from January 2016 to December 2017 within Bulgaria. Each node corresponds to an economic sector whereas the width of the directed edges represent the amount of VAT exchanged in the direction of the edge. The correspondence between the economic sectors and the capital letters used as node labels can be found in the supplementary material.

Network anomaly detection algorithms have already seen significant development during the last decade, see for example Akoglu et al. (2015); Baltoiu et al. (2019); Fernandes et al. (2019). For additional contributions, particularly in fraud detection in other settings, see Irofti et al. (2019); Elliott et al. (2019) who focus on breaking a network into communities to search for fraud within communities, deep learning based approaches, as well as a combination of spectral methods with motif based statistics to detect anomalous nodes.

VAT fraud, and MTIC fraud in particular, is achieved by exploiting the many-stages of invoice-transactions within B2B transactions involved in the export, import, and re-export of goods. Under MTIC fraud, fraudulent businesses import goods from overseas, VAT-free, before selling them on to domestic buyers, charging them VAT. This process often continues, with the goods being re-exported and re-imported for the fraud to continue. What this means, in practice, is that the fraud leaves a *pattern of transactions* between businesses which, naturally, takes the form of a weighted network. A typical network on which VAT applies is illustrated in Figure 1, which depicts VAT transactions across economic sectors in Bulgaria in 2016/2017.

When studying interactions we supplement the recordings of the interactions that are weighted and directed. We also have information on node–specific (estimated) vector with fraud probabilities based on node–specific covariates. VAT–fraud inevitably is a community activity. We therefore seek to determine communities whose members are likely to be suspected of VAT–fraud. Normally communities in networks are determined from the graph Laplacian, calculated from the (weighted) adjacency matrix. We now seek to combine our understanding of the node-specific structure of coherent interaction behaviour. This leads us to determine a *corrected version of the Laplacian* for determining anomalous communities, thus, recognizing the anatomy of fraud as a combination of individual propensity with community opportunity.

We update a vertex–specific binary vector using the singular value decomposition of the regularized graph Laplacian via estimated anomaly probabilities. This encapsulates both using information across nodes, as well as node-specific information, resulting in a probability that node i is involved in fraudulent activities. We show how to implement this method on a population–sized data set involving over $300,000$ entities, based on all VAT–registered businesses in Bulgaria. By implementing a design of a population–scale anomaly detection scheme, our work naturally fits into the program of population–scale inference in computational social science.

2 Data: VAT Administrative Network Data

To put our network anomaly detection algorithm into context, let us describe the characteristics of the data we analyze in this paper. VAT fraud, and particularly MTIC, is predominantly conducted through fictitious transactions and trading with the sole purpose of a cash outflow from the revenue authority of a country. The trader/taxable business at some point vanishes from the market without paying the due tax to the government. The objective of VAT fraudsters is, therefore, to conceal the fraud and go undetected using sophisticated transactions often involving many businesses across many sectors and countries, thus confounding any fraud detection attempts. MTIC fraud schemes involve organized criminals, missing or defaulting traders, buffer traders, broker traders, contra-traders end-customers (for acquisition fraud), freight forwarders and warehousing traders.

As a motivating data set, we deal with all VAT fraud within Bulgaria. Bulgaria is an early adopter of the VAT system, having introduced it in 1994, and became a member of the EU in 2007. As of 2017 the VAT Gap in Bulgaria (the difference between the tax revenue which should be collected and what is collected) was 12.2% just over the 11.5% EU average, so VAT fraud (a proportion of the VAT Gap) is considered sizeable (The European Commission, 2020). Developing, therefore, an early automatable identification system that identifies trades involved in the VAT fraud scheme is very valuable not only for the Bulgarian National Revenue Agency (BNRA) but also for all other revenue authorities facing the same problem. An important complementary feature is the identification of clusters consisted of taxpayers with similar characteristics. The cluster membership of each taxpayer can be also utilized by revenue authorities to deal efficiently with the complicated structure of large VAT networks and identify connections between VAT fraudsters.

We apply the developed algorithm on ledgers data for all $N = 312,762$ VAT registered taxpayers in Bulgaria in 2017. We conducted an out-of-sample exercise in which we trained the proposed model by constructing graphs that correspond to the monthly VAT returns filed by the taxpayers from January 2016 up to November 2017 and we aim to predict, probabilistically, the illegitimate taxpayers of December 2017. We compare our results with classification techniques that rely only on covariates that describe taxpayers' profiles without taking into account the network structure of the data. This out-of-sample exercise demonstrates that the graph information plays a key role in the efficient detection of anomalous vertices.

The constructed graphs are based on the aggregation of the VAT base from all transactions between each pair of VAT registered taxpayers. On average, 75% of the N VAT registered taxpayers conduct at least one transaction in a given month. Table

1 reports the composition of VAT base according to records required traders/taxable persons to declare under the Bulgarian VAT law as indicated in the first column of the table.[4] Tables 1 and 2 provide the summary statistics of the categories of transactions over the 24-month period. VAT fraud is potentially embedded within all aspects of VAT transactions described in Tables 1 and 2. As can be seen from the tables these transactions (in values and numbers) constitute a significant proportion of the total.

In addition to the aggregated VAT base of the invoices, the analysis utilizes a set of features that describe the profile of each VAT registered trader. These include, the size of the VAT registered company, the age of the company, labour costs as well as the classification of the transactions conducted by the registered traders. Importantly, each registered taxpayer has been classified as "high–risk" or "low–risk" by BNRA by utilizing past information of fraudulent activity as well as operational knowledge, this corresponding to $\mathbf{Y}$ (testing and training data) and $\check{\mathbf{Y}}$ (the training data alone). It is worth noting that the average proportion of "high–risk" traders/taxable persons during the time period is 1% per month. The value of goods/services and the corresponding VAT base that each trader has transacted with "high–risk" traders is also available and classified according to the categories displayed in Tables 1 and 2.

Table 1: The total VAT base reported on sells invoices and imports for the years 2016 and 2017 across the categories of VAT transactions. All values are expressed in local currency.

	2016	2017
Sum of VAT base (sells & imports)	305,386,748,486	334,040,088,090
ICA (%)	10.8	10.7
ICD (%)	9.3	9.4
9% (%)	0.7	0.6
Services from EU (%)	6.6	6.3
Deliveries from outside Bulgaria (%)	2.1	2.4
Exports to third countries (%)	6.9	7.3
Imports from third countries (%)	5.3	6.2
0% special deliveries (%)	0.1	0.1
TA (%)	0.6	0.7
TD (%)	0.8	0.8

3 Methods: Anomaly Detection

The main challenge in detecting network anomalies is the classification of the network features into "normal" and "anomalous". In a network there is more than one type of

---

[4]The following abbreviations have been used: Inter-community Acquisitions (ICA), Inter-community Deliveries (ICD), Triangural Acquisitions (TA), Triangular Deliveries (TD).

Table 2: Total number of transactions for the years 2016 and 2017 across the categories of VAT transactions. All values are expressed in local currency.

	Median	Minimum	Maximum
ICA	390,843	147,787	528,678
ICD	217,397	85,782	305,091
9%	193,019	91,969	251,009
Services from EU	212,614	64,362	284,953
Deliveries from out of Bulgarian territory	84,360	18,529	168,261
Exports to third countries	280,437	96,196	413,784
Imports from third countries	49,237	14,797	65,271
0% special deliveries	8,148	2,801	14,733
TA	12,480	2,619	16,244
TD	13,023	1,689	18,067

anomaly. If, additionally, one considers the constraint of a method then we must also take computational scalability into consideration.

A review of the issues in network anomaly detection appears in Akoglu et al. (2015) and Fernandes et al. (2019). Given the nature of MTIC fraud anomalous activities are characterised both by nodal covariates, and by groups of nodes having a particular pattern of association. The approach taken in the present paper merges these two points-of-view (groups and nodal perspectives).

We start from the adjacency matrix $\mathbf{A}(G)$ constructed from the (weighted) edges $E(G)$ of interactions between nodes (taxpayers) $V(G)$ with weights $\{w_{ij}\}$. The adjacency matrix thus takes the form of

$$a_{ij}(G) = \begin{cases} w_{ij} & \text{if} & (i,j) \in E(G) & i,j \in \{1, \dots N\} \\ 0 & & \text{otherwise} \end{cases}.$$

If the w_{ij} are not symmetric then we define $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{A}^T$ otherwise we take $\widetilde{\mathbf{A}} = \mathbf{A}$. Figure 1 shows the sector–specific network of tax interactions from 2016 and 2017.

The proposed method employs the network $\mathbf{A}$, coupled with node-specific covariates $\mathbf{X}$, to estimate probabilities of suspected fraud $\widetilde{\mathbf{p}}$. The approach begins by estimating probabilities of fraud for each individual node i without using the network structure and then proceeds with understanding of how our knowledge of the group structure of G enriches our initial estimates providing more accurate predictive probabilities of a fraudulent node.

3.1 Initial Detection of Anomalous Nodes

Our data set contains weighted interactions as well as other variables (covariates) indexed by the nodal number $i = 1, \ldots, N$, as we assume that there are $N \in \mathbb{N}$ nodes in the network. We collect the node-specific covariates in the $N \times p$ matrix $\mathbf{X}$ where p is the number of available covariates. To classify the nodes without initially using their network community structure we shall use the $\mathbf{X}$ to implement binary classification using the scalable XGBoost method (Chen and Guestrin, 2016); see the supplementary material for a brief description of the XGboost method while gradient boosting is further detailed in James et al. (2013). To be able to implement this method, we assume availability of training data $\check{\mathbf{Y}}, \check{\mathbf{X}}$ with previously identified cases of fraud, versus cases of not detected fraud, with associated covariates. Thus, after training the XGboost algorithm, we obtain an output $\widehat{\mathbf{p}}(\mathbf{X}) \equiv \widehat{\mathbf{p}}$. This has not taken the network structure of fraudulent activities into account, but depends on the nodal characteristics via $\mathbf{X}$.

3.2 Incorporating Graph Characteristics of VAT Fraud

VAT fraud is predominantly a community or group activity and cannot normally be implemented simply by a single actor. To be able to model the group structure of activities, we need to detect those groups or communities that are more probable to be involved in fraudulent behaviour. This can be done by fitting a group model that identifies the nodes present in any group. This fit can either be implemented under the assumption that there are true blocks in the data (Newman, 2012) or just a propensity of a range of nodes to behave like a grouping (Olhede and Wolfe, 2014).

The most common method to extract community structure from a network is *spectral clustering* (Chung and Graham, 1997) which is based on a spectral partition from the graph Laplacian matrix with $\mathbf{D} = \text{diag}\{d_1, \ldots, d_N\}$ as a diagonal matrix and so

$$\mathbf{L}(\alpha, \tau) = \mathbf{D}_\tau^{-1/2} \widetilde{\mathbf{A}} \mathbf{D}_\tau^{-1/2} + \alpha \widehat{\mathbf{p}} \widehat{\mathbf{p}}^T, \ \mathbf{D}_\tau = \mathbf{D} + \tau \mathbf{I}. \tag{1}$$

Note that $\widetilde{\mathbf{A}}$ is symmetric. There are a number of possible "Laplacians" that we might have defined, both in terms of the Laplacian, the graph Laplacian (Chung and Graham, 1997), and various regularized Laplacians (Chaudhuri et al., 2012; Qin and Rohe, 2013; Binkiewicz et al., 2017). There has also been some debate about using Laplacian spectral clustering or adjacency spectral clustering (Priebe et al., 2019). The blockmodel for weighted adjacency matrices has been discussed in detail in Peixoto (2018). We choose to adopt adjacency spectral clustering, combined with a choice of regularization that depends on the two tuning parameters τ and α.

We can see directly from (1) that if the adjacency matrix is zero and there were

no network structure in the data then we would only cluster on the values of the vector $\widehat{\mathbf{p}}$. If, on the other hand, τ was chosen to be zero, then there would be no regularization when inverting the degree matrix. As a final step we wish to update $\widehat{\mathbf{p}}$ to $\widetilde{\mathbf{p}}$ using the graph structure and then to select a threshold to determine which $\widetilde{\mathbf{p}}$ are large enough to warrant further investigation. This is achieved again via XGboost and by utilizing the loadings of the eigenvectors of $\mathbf{L}(\alpha, \tau)$; see in the supplementary material for more details.

4   Population–Scale MTIC Fraud Detection

4.1   VAT Fraud Detection Example

The aim of VAT fraud detection is to determine which taxpayers are suspected as being potential actors in a fraud scheme. We therefore apply the proposed algorithm, Graph Informed Multiscale Anomaly Detector (GIMAD)[5], to the data presented in Section 2 corresponding to VAT returns from the years 2016-2017 for $312,762$ taxpayers. Taxpayers submit monthly VAT returns, implying that we have 24 temporal snapshots which we will train on 23 months of transactions and then validate on 1 month of transactions. An edge in the network implies that there is at least one tax transaction between the two taxpayers.

Prediction of probabilities of risky taxpayers is achieved by first training the XG-boost algorithm with inputs a binary response vector $\breve{\mathbf{Y}}$ and the $N \times p$ matrix $\breve{\mathbf{X}}$ consisted of the available covariates which are: the sector in which each taxpayer (company) is registered, the number of the employees, the size and the labour costs of each company and other records that taxpayers declare with their VAT returns. In particular, for each taxpayer and for each category in the first column of Table 1 we compute the number of transactions that they have conducted along with the corresponding VAT base and the proportion of these transactions conducted with "high–risk" taxpayers as indicated by the input vector $\breve{\mathbf{Y}}$. We note that these type of covariates are a subset of the risk-based criteria which the Bulgarian authorities employ in order to prioritize the taxpayers with respect to their riskiness of being involved in an MTIC fraud. We also construct covariates by utilizing the characteristics of the 23 observed graphs. We calculate for each vertex its mean (across the observed graphs) degree, strength and centrality. The resulting matrix has $p = 49$ columns. Then, we utilize the $N \times p$ matrix $\mathbf{X}$ consisted of the covariates that correspond to the month (December 2017) that we wish to predict risky probabilities in order to obtain the vector $\widehat{\mathbf{p}}$ appearing in (1).

---

[5]See in the supplementary material for the algorithmic description of the steps of the proposed GIMAD.

The input adjacency matrix $\mathbf{A}$ corresponds to the adjacency matrix of a directed weighted graph, constructed by the VAT returns submitted in December 2017. In our case $\mathbf{A}$ is an asymmetric matrix so we construct a symmetric matrix $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{A}^T$. The undirected graph whose adjacency matrix is $\widetilde{\mathbf{A}}$ has the same edges as the original graph, but directed edges have been replaced with undirected edges with a sum of the weights associated with the edge in question. Community detection methods that are based on $\widetilde{\mathbf{A}}$ tend to group nodes that share similar incoming <u>and</u> outgoing edges (Satuluri and Parthasarathy, 2011). We find this symmetrization reasonable as VAT-registered traders that perform fraudulent activity, it is reasonable to assume, have common trading patterns. Finally, we do need to determine tuning parameters with this method, e.g. both α and τ. We follow the advice of Qin and Rohe (2013) and set $\hat{\tau} = N^{-1} \sum_{i=1}^n d_{ii} = \bar{d}$, the average degree. The value of α can be determined from the eigenvectors of $\mathbf{D}_\tau^{-1/2} \widetilde{\mathbf{A}} \mathbf{D}_\tau^{-1/2}$ and $\widehat{\mathbf{p}}$. See for example in Binkiewicz et al. (2017) where the authors show how to set α such that the information contained in $\mathbf{D}_\tau^{-1/2} \widetilde{\mathbf{A}} \mathbf{D}_\tau^{-1/2}$ as well as in $\widehat{\mathbf{p}}$ is captured in the leading eigenspace of $\mathbf{L}(\alpha, \hat{\tau})$.

4.2 Out–of–Sample Detection

To test the performance of our anomaly detection algorithm we have designed an out–of–sample detection exercise. We construct a time series of graphs from the 24 months of observations corresponding to the monthly data of 2016 and 2017.

Our first step in classifying the 24th month of observations from the other 23 corresponds to a binary vector that indicates the anomalous vertices of "high–risk" taxpayers, a matrix of covariates and an adjacency matrix. The binary vector $\breve{\mathbf{Y}}$ that we input is a classification of "high–risk" and "low–risk" taxpayers, as calculated by BNRA up to the point November 2017. We note that this is an unbalanced classification problem as the proportion of fraudulent node is unlikely to be as large as one half (Hand and Vinciotti, 2003). This corresponds to assigning a different loss to the different types of miss-classification. To deal with this class imbalance problem we apply the method of random oversampling by randomly re-sampling the set of "high–risk" taxpayers in order to construct a balanced dataset. We have chosen the technique of oversampling among others in order to keep the proposed method simple without loosing any information carried on the original data; see, for example, Menon et al. (2013) for a comparison of the several techniques that have been developed to deal with data imbalance problems. To carry out the out–of–sample analysis we use the weighted directed graph made from the VAT returns submitted in December 2017. The value of the tuning parameter α should be chosen to balance the network structure, as captured by $\widetilde{\mathbf{A}}$, versus the individual probabilities of $\widehat{\mathbf{p}}$. We implemented sensitivity analysis to determine a value of 0.01 for
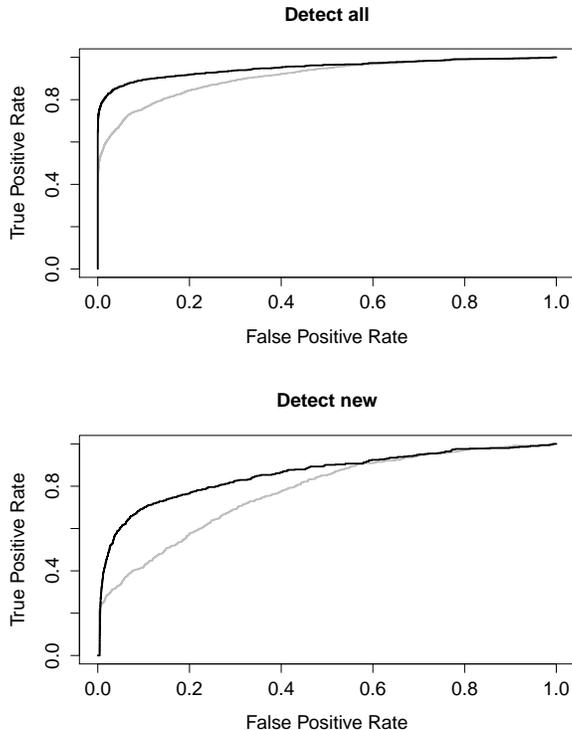
Figure 2: ROC curves that compare the out-of-sample classification performance of the proposed method (black line) with the out-of-sample classification conducted without utilising the network information (gray line). Top: The comparison corresponds to the detection of all the "high-risk" taxpayers of December 2017. Bottom: The comparison corresponds to the detection of taxpayers that entered the risk registration list of BNRA at December 2017.

this parameter. To calculate the spectral decomposition of the matrix $\boldsymbol{L}(0.01, \hat{\tau})$ in (1) we employed the Lanczos bidiagonalization method (Baglama and Reichel, 2005) and we stopped the algorithm after computing the first $K = 200$ eigenvalues and eigenvectors by noting that after that value the eigenvalues were quite similar; see Figure 1 in the supplementary material for their values. In the supplementary material we also present Algorithm 1 which summarizes the steps of the proposed method. The application of Algorithm 1 on the described dataset required almost 3 hours, on a Laptop with a 1.6 GHz Dual-Core Intel Core i5 CPU running R 4.0.0 R Core Team (2021).

4.3 Determining the Accuracy of the Proposed Method

We evaluate the proposed technique by trying to predict the provided list of risky taxpayers, as occurring in December 2017. We can observe directly from the list of risky taxpayers that 64% of the high–risk registrations of taxpayers in December 2017 had in fact been determined as "high–risk" already in November 2017. The remaining 36% were registered for the first time as "high–risk" in December 2017. We, therefore, address the

two tasks of i) predict all risky registrations in 2017, and ii) predict only the new risky registrations in 2017.

To determine the performance of our novel methodology we compare the receiver operating characteristic (ROC) curves (Hsieh et al., 1996) produced by our method and by using XGboost classification without the network information. Figure 2 illustrates that our algorithm outperforms the simple XGboost algorithm in both the old and new taxpayers in December 2017. This provides strong evidence of the added utility to combine both individual and group patterns to detect fraud.

4.4   Policy Evaluation of the Algorithmic Output

The policy gain of the automated detection algorithm proposed in this paper is clear;[6] currently BNRA applies risk–based rules on all the submitted tax returns and monthly prioritizes 15,000 of these returns as "high–risk". By implementing further selection criteria those 15,000 are whittled down to 500, and finally via auditing 100 taxpayers are identified as having been part of VAT fraud. The method proposed in this contribution provides a fully automated mechanism for identifying VAT fraudsters. Automation has a number of clearly identified advantages, reducing cost, increasing transparency and reproducibility, explicitly balancing the information obtained from a single taxpayer versus that provided by the population–scale data. The out-of-sample exercise shows a clear benefit in identification for a fixed false positive rate. In particular, our contribution to policy is identified as having determined 200 taxpayers with the highest fraud probabilities, where 100 of them entered the risky list for the first time in December 2017. By automation we have reduced the set of 500 identified by the BNRA selection procedure that relied on a human–implemented selection procedure.

Finally, Figure 3 displays the number of new entries in the risky list that we can identify for a given number of taxpayers reported as suspicious. The figure indicates that reducing 200 to 50 reported taxpayers we minimize out false positive rate since 40 of them entered the risky list of BNRA at December 2017 indeed. Allowing for more false positives[7] and increasing 200 to 500 which is the number currently audited by BNRA we can predict more than 120 "high–risk" taxpayers while we can find more than 140 by reporting $2,000$ taxpayers to be audited.

---

[6]Though the method is applied on data from BNRA its application is, of course, broader and can be employed by other revenue authorities as well.

[7]As false positives we denote taxpayers that have not been audited by BNRA with respect to participation in VAT fraud scheme.
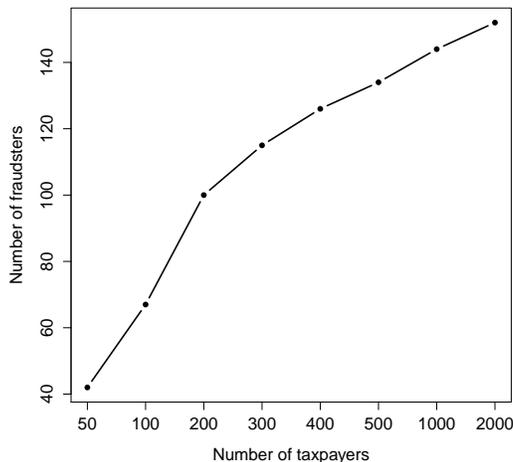
Figure 3: The X-axis indicates the number of taxpayers that we need to report in order to identify the number of taxpayers that have entered the risk–list of BNRA for a first time in December 2017 (Y-axis).

5 Discussion

When determining anomalous events the basic question we must answer is: Are we willing to define what is "normal" or what is "abnormal"? Yet, a problem is always that we observe more normal data than abnormal, and that abnormal features are quite rare (Donoho and Jin, 2008). In very large data sets, such as the VAT system we analyzed, and consisted of 312,762 taxpayers, we would still like to borrow information across taxpayers to reduce variability in characterising individual taxpayers and help our decision-making. We are therefore going from studying the large-scale set of normal individuals to the micro–scale set of abnormal individuals. The idea in this paper has been to combine our understanding of the micro–scale of individuals with the meso–scale of local segments of the population engaged in suspicious activities in order to reduce variability in the delivered predictions. Of course detecting anomalies is just an admissible strategy in response to a particular choice of regulatory policy (Black and Baldwin, 2012); over time policy may change necessitating a change in the algorithm.

Is then the notion that individual variability should be updated with medium–scale structure present in the data unique to VAT and tax fraud? Other authors studying fraud in money–laundering have identified the importance of community, if combined with dictionary learning (Baltoiu et al., 2019). Additionally, in bioinformatics, in particular metabolomics, interactions measures are used to mediate individual loci exceptionality (Shin et al., 2014). The notion of a group model combines naturally with distributional local permutation invariance (Kallenberg, 2006), allowing us to locally average, but still combining group structure with individual exceptionality by the prob-

ability estimate that is moderated by the group structure. This can be considered for networks as a generalization of the degree–corrected stochastic blockmodel, except here, instead of moderating by a proclivity to connect of the network, we moderate by the exceptionality of the individual taxpayer. This allows us to characterise how our individual perception of exceptionality is not just a combination of "normal" (population–scale), "abnormal" (micro–scale), and segments of the population that exhibit similar characteristics (meso–scale). Our understanding of populations will inevitably be patchy; but needs to be glued together from our understanding of variability at different scales.

References

Akoglu, L., H. Tong, and D. Koutra (2015). Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery 29*(3), 626–688.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science 355*(6324), 483–485.

Baglama, J. and L. Reichel (2005). Augmented implicitly restarted lanczos bidiagonalization methods. *SIAM Journal on Scientific Computing 27*(1), 19–42.

Baltoiu, A., A. Patrascu, and P. Irofti (2019). Community-level anomaly detection for anti-money laundering. *arXiv preprint arXiv:1910.11313*.

Binkiewicz, N., J. T. Vogelstein, and K. Rohe (2017). Covariate-assisted spectral clustering. *Biometrika 104*(2), 361–377.

Black, J. and R. Baldwin (2012). When risk-based regulation aims low: A strategic framework. *Regulation & Governance 6*(2), 131–148.

Chaudhuri, K., F. Chung, and A. Tsiatas (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. In *Conference on Learning Theory*, pp. 35–1.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Chung, F. R. and F. C. Graham (1997). *Spectral graph theory*. Number 92. American Mathematical Soc.

Donoho, D. and J. Jin (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences 105*(39), 14790–14795.

Ebrill, M. L. P., M. M. Keen, and M. V. P. Perry (2001). *The modern VAT*. International Monetary Fund.

Elliott, A., M. Cucuringu, M. M. Luaces, P. Reidy, and G. Reinert (2019). Anomaly detection in networks with application to financial transaction networks. *arXiv preprint arXiv:1901.00402*.

Fernandes, G., J. J. Rodrigues, L. F. Carvalho, J. F. Al-Muhtadi, and M. L. Proença (2019). A comprehensive survey on network anomaly detection. *Telecommunication Systems 70*(3), 447–489.

Hand, D. J. and V. Vinciotti (2003). Choosing k for two-class nearest neighbour classifiers with unbalanced classes. *Pattern recognition letters 24*(9-10), 1555–1562.

Hsieh, F., B. W. Turnbull, et al. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Annals of statistics 24*(1), 25–40.

Irofti, P., A. Patrascu, and A. Baltoiu (2019). Fraud detection in networks: State-of-the-art. *arXiv preprint arXiv:1910.11299*.

Jackson, M. O. and A. Wolinsky (1996). A strategic model of social and economic networks. *Journal of economic theory 71*(1), 44–74.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learning*, Volume 112. Springer.

Kallenberg, O. (2006). *Probabilistic symmetries and invariance principles*. Springer Science & Business Media.

Keen, M. and B. Lockwood (2010). The value added tax: Its causes and consequences. *Journal of Development Economics 92*(2), 138–151.

Keen, M. and S. Smith (2006). Vat fraud and evasion: What do we know and what can be done? *National Tax Journal 51*, 861–887.

Lazer, D., A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. (2009). Life in the network: the coming age of computational social science. *Science (New York, NY) 323*(5915), 721.

Lazer, D. M., A. Pentland, D. J. Watts, S. Aral, S. Athey, N. Contractor, D. Freelon, S. Gonzalez-Bailon, G. King, H. Margetts, et al. (2020). Computational social science: Obstacles and opportunities. *Science 369*(6507), 1060–1062.

Margetts, H. and C. Dorobantu (2019). Rethink government with ai.

Menon, A., H. Narasimhan, S. Agarwal, and S. Chawla (2013). On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning*, pp. 603–611.

Newman, M. E. (2012). Communities, modules and large-scale structure in networks. *Nature physics 8*(1), 25–31.

Olhede, S. C. and P. J. Wolfe (2014). Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences 111*(41), 14722–14727.

Peixoto, T. P. (2018). Nonparametric weighted stochastic block models. *Physical Review E 97*(1), 012306.

Priebe, C. E., Y. Park, J. T. Vogelstein, J. M. Conroy, V. Lyzinski, M. Tang, A. Athreya, J. Cape, and E. Bridgeford (2019). On a two-truths phenomenon in spectral graph clustering. *Proceedings of the National Academy of Sciences 116*(13), 5995–6000.

Qin, T. and K. Rohe (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in neural information processing systems*, pp. 3120–3128.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Satuluri, V. and S. Parthasarathy (2011). Symmetrizations for clustering directed graphs. In *Proceedings of the 14th International Conference on Extending Database Technology*, pp. 343–354.

Shin, S.-Y., E. B. Fauman, A.-K. Petersen, J. Krumsiek, R. Santos, J. Huang, M. Arnold, I. Erte, V. Forgetta, T.-P. Yang, et al. (2014). An atlas of genetic influences on human blood metabolites. *Nature genetics 46*(6), 543–550.

The European Commission (2020). Reports on the vat gap in the eu-28 member states.

SUPPLEMENTARY MATERIAL

DETECTING ANOMALIES IN HETEROGENEOUS POPULATION–SCALE VAT NETWORKS

Angelos Alexopoulos[a,1], Petros Dellaportas[b,c,d], Stanley Gyoshev[ℓ], Christos Kotsogiannis[e,f,k], Sofia C. Olhede[b,g], Trifon Pavkov[ℓ,h]

[a]MRC Biostatistics Unit, University of Cambridge, University Forvie Site, Robinson Way, Cambridge CB2 0SR, UK. Email: angelos@mrc-bsu.cam.ac.uk.
[b]Department of Statistical Science, University College, Gower Street, London, WC1E 6BT, UK. Email: p.dellaportas@ucl.ac.uk.
[c]Department of Statistics, Athens University of Economics and Business, Greece.
[d]The Alan Turing Institute, London, UK.
[e]Department of Economics, University of Exeter Business School, Streatham Court, Rennes Drive, Exeter EX4 4PU, England, UK. Email: c.kotsogiannis@exeter.ac.uk.
[f]ESIfo, Munich, Germany.
[g]Institute of Mathematics, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland. Email: sofia.olhede@epfl.ch.
[h]National Revenue Agency, Sofia, Bulgaria. Email: t.pavkov@nra.bg.
[k]Tax Administration Research Centre (TARC), University of Exeter.
[ℓ]Department of Finance, University of Exeter Business School, Streatham Court, Rennes Drive, EX4 4PU, England, UK. Email: s.gyoshev@exeter.ac.uk.

1   Introduction

This document supplements the material presented in the manuscript "Detecting anomalies in heterogeneous population–scale VAT networks". Section 2 presents the codes of the economic sectors displayed by Figure 1 of the main paper. Section 3 provides a brief description of the XGboost algoritm by Chen and Guestrin (2016) and explains how it is employed in the proposed anomaly detection method. Section 4 presents Algorithm 1 which summarizes the steps of the developed anomaly detection technique. Finally, Section 5 shows additional results from the eigendecomposition of the covariate assisted graph Laplacian defined by equation (1) in the main paper.

2   Table with Economic Sector Codes

Table 1 displays the codes of the economic sectors in Bulgaria classified according to the Nomenclature of Economic Activities (NACE) system.

Table 1: Sector codes according to the Nomenclature of Economic Activities (NACE) classification system.

Code	Sector
A	Agriculture, forestry and fishing
B	Mining and quarrying
C	Manufacturing
D	Electricity, gas, steam and air conditioning supply
E	Water supply; sewerage; waste management and remediation activities
F	Construction
G	Wholesale and retail trade; repair of motor vehicles and motorcycles
H	Transporting and storage
I	Accommodation and food service activities
J	Information and communication
K	Financial and insurance activities
L	Real estate activities
M	Professional, scientific and technical activities
N	Administrative and support service activities
O	Public administration and defence; compulsory social security
P	Education
Q	Human health and social work activities
R	Arts, entertainment and recreation
S	Other services activities
NA	Not available information of the economic activity

3   Brief Description of the XGboost Algorithm

For a dataset in which $\boldsymbol{Y}$ is an N-dimensional vector with responses and $\boldsymbol{X}$ is an $N \times p$ matrix with features (covariates) XGboost has originally described by Chen and Guestrin (2016) as an ensemble of S regression trees where a prediction $\hat{y}_i$ is obtained as

$$\hat{y}_i = \sum_{s=1}^{S} f_s(\boldsymbol{X}_i), \ \ f_s \in \mathcal{F}, \tag{1}$$

where $\mathcal{F}$ is the space of the regression trees and each function f_s corresponds to an independent tree structure with T number of leaves and leaf scores $\boldsymbol{V} \in \mathbf{R}^T$. We learn

the functions f_s by minimizing the regularized objective

$$\mathcal{L} = \sum_{i=1}^{N} \ell(\hat{y}_i, y_i) + \sum_{s=1}^{S} \Omega(f_k), \qquad (2)$$

where ℓ is a differentiable convex loss function which measures the difference between the predicted and the target label and

$$\Omega(f) = T\gamma + (1/2)\lambda \sum_{j=1}^{T} v_j^2,$$

is a regularization term which penalizes the complexity of the model to avoid over-fitting, T is the number of leaves and v_j the score on the jth leaf while γ and λ are constants to control the degree of regularization. Since the parameters of the model in (2) are the functions f_s traditional optimization methods of the Euclidean space cannot be used. Instead the model is trained in an additive manner by first noting that the additive structure of the prediction in (1) implies that $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(\boldsymbol{X}_i)$, where the superscript t denotes the tth iteration of the optimization procedure. Then, objective in (2) becomes

$$\mathcal{L}^{(t)} = \sum_{i=1}^{N} \ell\big(y_i, \hat{y}_i^{(t-1)} + f_t(\boldsymbol{X}_i)\big) + \Omega(f_t). \qquad (3)$$

After a second order Taylor approximation and by removing all the constant terms we have that

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{N} [g_i f_t(\boldsymbol{X}_i) + \frac{1}{2} h_i f_t^2(\boldsymbol{X}_i)] + \Omega(f_t), \qquad (4)$$

where $g_i = \partial_{\hat{y}^{(t-1)}} \ell(y_i, \hat{y}^{(t-1)})$ and $h_i = \partial_{\hat{y}^{(t-1)}}^2 \ell(y_i, \hat{y}^{(t-1)})$ are first and second order gradient of the loss function. By expanding the regularization term Ω and noting the quadratic form of equation (4) it is easy to find the optimal weights $\boldsymbol{V}$. Therefore, for a given tree structure we can compute the optimal leaf weights $\boldsymbol{V}$ and to calculate the corresponding value of (3). Since it is impossible to make these calculations for all the possible tree structures, Chen and Guestrin (2016) show that for the loss function in (3) it is straightforward to calculate a score for a leaf node during splitting and based on this score they propose to utilize the so-called exact greedy algorithm in order to detect the split point that results in maximum loss reduction. Therefore, XGboost becomes a scalable method which is more than ten times faster than other popular algorithms in a wide range of problems (Chen and Guestrin, 2016).

In the anomaly detection method that we build we utilize the XGboost algorithm twice in order to estimate node-specific anomalous probabilities. We have denoted the training binary responses, which we give as input to XGboost, with $\check{\mathbf{Y}}$. As described in

Section 4.A of the main paper in order to perform an initial estimation of anomalous, node-specific, probabilities we train the XGboost algorithm by using the covariates in $\check{\mathbf{X}}$ and then we use the covariates in $\mathbf{X}$ to obtain the vector $\widehat{\mathbf{p}}$ appearing in the matrix $\mathbf{L}(\alpha, \tau)$ defined by equation (1) of the main paper. In the final step of our method we use as covariates the loadings of the eigenvectors of the matrix $\mathbf{L}(\alpha, \tau)$ in order to update, via XGboost again, $\widehat{\mathbf{p}}$ to the graph-informed anomalous probabilities $\widetilde{\mathbf{p}}$. The function ℓ in (2) is the negative logarithm of the Bernoulli probability mass function.

4 The Proposed Graph Informed Multiscale Detector (GIMAD)

Algorithm 1 summarizes the steps of the developed network anomaly detection technique. The proposed algorithm requires as inputs the graph structure (network adjacency matrix) of the data as well as a vertex specific set of covariates and binary indicators of vertex anomalousness. The output of the algorithm is consisted of a vector with estimated anomaly probabilities for each vertex and a vector of cluster memberships for the vertices. We note that the input adjacency matrix can be either symmetric (undirected graph) or non-symmetric (directed graph) since in the latter case sophisticated techniques for transforming a directed graph to a non-directed one have been proposed in the recent literature and can be easily employed; see, for example, Malliaros and Vazirgiannis (2013) for a detailed discussion.

Algorithm 1 GIMAD

Input: $N \times N$ network adjacency matrix $\boldsymbol{A}$; N-dimensional vertex specific binary vector $\boldsymbol{Y}$; $N \times p$ matrix $\boldsymbol{X}$ with vertex specific covariates; tuning constant $\alpha > 0$; positive integer K.

1: **if** $\boldsymbol{A}$ symmetric **then**
2:     Set $\tilde{\boldsymbol{A}} = \boldsymbol{A}$
3: **else**
4:     Set $\tilde{\boldsymbol{A}}$ to be the symmetric matrix obtained after suitable transformation on $\boldsymbol{A}$.
5: **end if**
6: Predict anomaly probabilities $\widehat{\mathbf{p}}$ by first training XGboost[1] on responses $\boldsymbol{Y}$ and covariates $\boldsymbol{X}$.
7: Calculate $\boldsymbol{L}(\alpha, \hat{\tau})$ defined by equation (1) in the main paper.
8: Compute the eigendecomposition $\boldsymbol{L}(\alpha, \hat{\tau})$ and form the $N \times K$ matrix $\boldsymbol{U}$ with columns the eigenvectors that correspond to the K largest eigenvalues.
9: Normalize each row in $\boldsymbol{U}$ to have unit length and form the $N \times K$ matrix $\boldsymbol{W}$ with $w_{ik} = u_{ik}\sqrt{\lambda_k}$.
10: Estimate anomaly probabilities $\widetilde{\mathbf{p}}$ by using XGboost with responses $\boldsymbol{Y}$ and features $\boldsymbol{W}$.
11: Treat each normalized row of $\boldsymbol{U}$ as point in $\mathbb{R}^K$ and run a k-means clustering algorithm with K clusters; if the ith row of $\boldsymbol{U}$ falls in the kth cluster assign node i to cluster k.

Output: N-dimensional vector $\widetilde{\mathbf{p}}$ with vertex specific anomaly probabilities; N-dimensional vector $\boldsymbol{C}$ with vertex specific cluster memberships.

---

5   Results from the Spectral Decomposition

Figure 1 displays the first 200 eigenvalues of the matrix $\boldsymbol{L}(0.01, \hat{\tau})$ computed by using the Lanczos bidiagonalization algorithm (Baglama and Reichel, 2005). Figure 2 presents the mean of each loading vector separately for the "low–risk" taxpayers, for the "high–risk" taxpayers that we used to train GIMAD and for the "high–risk" taxpayers that we aimed detect. It is clearly indicated by the Figure that for the "high–risk" taxpayers exists one eigenvector for which the mean of its loadings is much higher than the means of the loadings that correspond to the rest eigenvectors. By noting that in the case of "low–risk" taxpayers the mean loadings for all the eigenvectors have similar values we conclude that using, in the 9th step of Algorithm 1, the columns of matrix $\boldsymbol{W}$ as

---
[1]We utilize the r-package `xgboost` (Chen et al., 2019).

features in a XGboost algorithm we obtain an accurate classification between "high-"
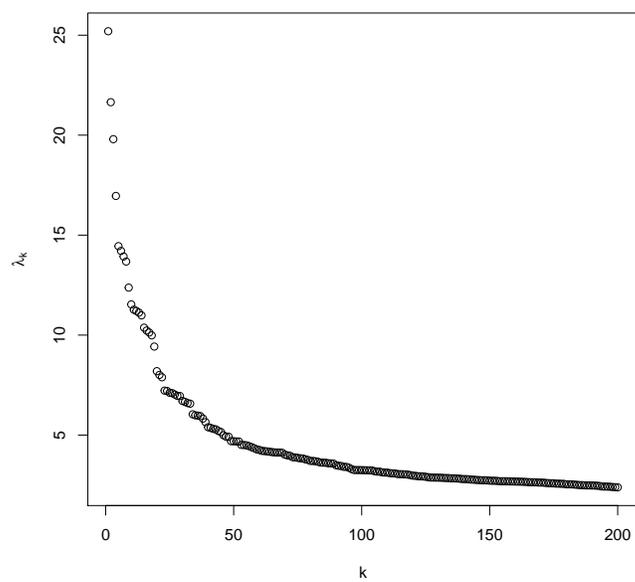and "low–risk" taxpayers.

Figure 1: The first $K = 200$ eigenvalues of the matrix $\boldsymbol{L}(0.01, \hat{\tau})$ computed by using the
Lanczos bidiagonalization algorithm (Baglama and Reichel, 2005).
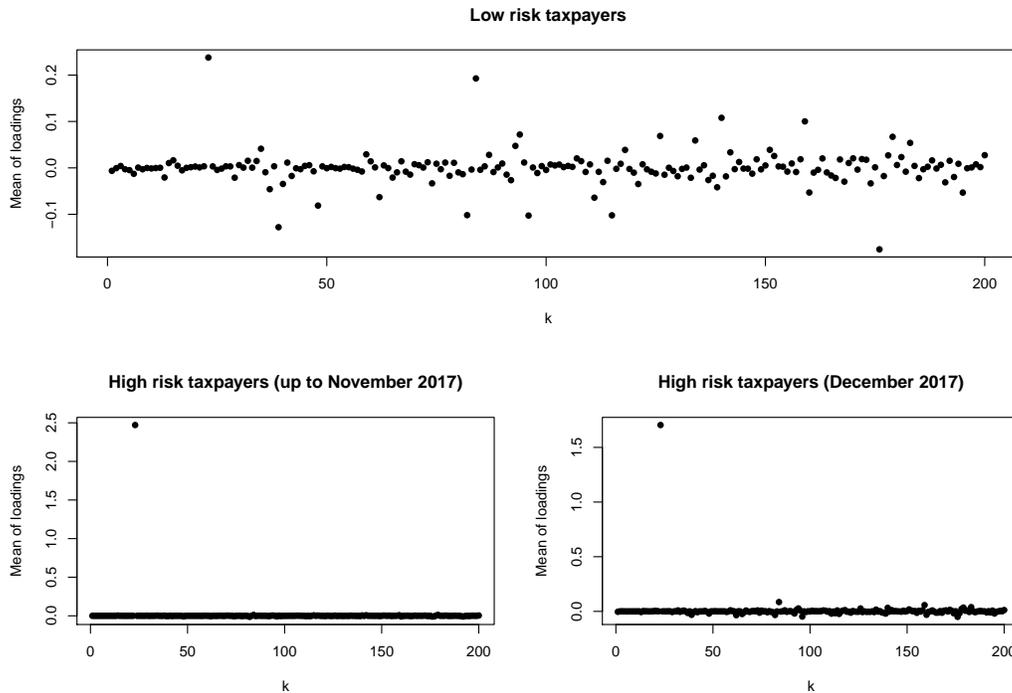
Figure 2: Mean of the loadings that correspond to the first $K = 200$ eigenvalues. The "high–risk" taxpayers are separated to those that we used to train our method and those that we aimed to detect. The x-axis indicates the loading that corresponds to the kth eigenvalue.

References

Baglama, J. and L. Reichel (2005). Augmented implicitly restarted lanczos bidiagonalization methods. *SIAM Journal on Scientific Computing 27*(1), 19–42.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, and Y. Li (2019). *xgboost: Extreme Gradient Boosting.* R package version 0.90.0.2.

Malliaros, F. D. and M. Vazirgiannis (2013). Clustering and community detection in directed networks: A survey. *Physics Reports 533*(4), 95–142.