Detecting network anomalies in the Value Added Taxes (VAT) system

A TARC Policy Analysis Report

Angelos Alexopoulos, Petros Dellaportas, Stanley Gyoshev, Christos Kotsogiannis, Trifon Pavkov

A collaboration between TARC, UCL and the Bulgarian National Revenue Agency

This research has been funded by HSBC -- Alan Turing Institute under grant TEDSA2/100056

Detecting network anomalies in the Value Added Taxes (VAT) system

Angelos Alexopoulos[a], Petros Dellaportas[b], Stanley Gyoshev[c], Christos Kotsogiannis[d], Trifon Pavkov[e]

May 2020

Abstract: This report makes a methodological contribution regarding fraudulent detection in networks and applies it to high quality administrative Value Added Tax data.

Keywords: Anomalies detection; Networks; VAT networks; Carousel fraud; MTIC (Missing Trade Intercommunity).

Acknowledgments: We are grateful to the Bulgarian National Revenue Agency for outstanding collaboration. Financial support from HSBC-Alan Tuning Institute under grant TEDSA2/100056 is gratefully acknowledged. Without implicating, we also thank Sofia Olhede and Petya Staneva for comments and suggestions throughout the research. The research has been subject to a confidentiality agreement and no taxpayer data has been disclosed to the project team.

[a,d]Tax Administration Research Center (TARC), Department of Economics, University of Exeter Business School, Streatham Court, Rennes Drive, EX4 4PU, England, UK.

[c]Department of Finance, University of Exeter Business School, Streatham Court, Rennes Drive, EX4 4PU, England, UK.

[d]CESIfo, Munich, Germany.

[b]University College London and Athens University of Economics and Business.

[e]Bulgarian National Revenue Agency.

Executive Summary

VAT fraud is a significant problem, costing Revenue Agencies billions in lost tax revenues. To combat the problem the use of Data Analytics and Machine Learning approaches are needed.

This report:

- Makes a methodological contribution regarding fraudulent detection in networks, and
- applies the developed methodology to high quality administrative data on Bulgarian VAT transactions provided by the Bulgarian National Revenue Agency.

The key results are:

- The methodology has identified more than 90% of the high risk VAT-registered traders/taxable persons the National Revenue Agency has confirmed as such, and
- has identified 12 high risk VAT-registered traders/taxable persons out of a sample of 35 drawn from 8000 not identified as high risk by the National Revenue Agency.

Conclusion:

- The potential of the method developed is significant in identifying fraudulent transactions and close the VAT Gap stemming from such transactions.

1 Introduction

The objectives of this research project are twofold: Firstly, to make a contribution in 'detection anomalies' through the development of a methodology suitable for detecting fraudulent behaviours, and, secondly, to apply this to a high quality administrative data set of economic transactions which form the Value Added Taxes (VAT) base.

A Value Added Tax (in Bulgaria, in the EU and elsewhere) is typically levied on the invoice-credit basis, where the net tax liability of a business is calculated by subtracting from their sales the aggregate value of VAT paid on invoices for the inputs used in production. The VAT system therefore taxes sales of all goods and services but allows VAT-registered traders to deduct any VAT paid on purchases for business purposes, whether for resale or as an input into production.[1]

The two key features of the VAT system are the zero-rating of exports and the system of deferred payments.[2] Under deferred payment, VAT on imports into Bulgaria is levied not at the border but at the time of the importers next periodic VAT return. The consequence of this is that there may be a considerable time lag between the date at which the importing firm imports the goods and the time at which the National Revenue Agency seeks payment of the VAT due. This constitutes the VAT system's Achilles heel as it creates opportunities for VAT fraud.

VAT fraud is predominantly conducted through fictitious transactions and trading with the sole purpose of a cash outflow from the National Revenue Agency. This is achieved by exploiting the many-stages invoice-transactions between firms and involving a chain of cooperating firms across borders involved in the export, import, and re-export of goods. Fraudulent firms import goods from overseas, VAT-free, before selling them on to domestic buyers, charging them VAT. This process quite often continues, with the goods being exported and re-imported for the fraud to continue (a fraud which has been termed 'carousel fraud'). The trader/taxable person then at some point vanish 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 traders across many sectors and countries. VAT fraud is sophisticated and involves organized criminals, missing or defaulting traders, buffer traders, broker traders, contra traders end-customers (for acquisition fraud), freight forwarders, warehousing traders.[3] It involves mostly tangible commodities (typically of high value, low volume goods) but also intangible commodities-services. The Bulgarian National Revenue Agency takes significant measures against VAT fraud but unavoidably some escapes detection.

This reports shows that Data Analytics and Machine Learning can be significant instruments in the effort of the National Revenue Agency to collect the revenues due and combat VAT fraud. Building capacity along this dimension should be part of the strategic priority of the Agency.

[1] Bulgaria first brought in Value Added Tax (VAT) in 1994. Legislation for VAT in Bulgaria is contained within its Value Added Tax Act 2006 (Zakon za Danak varhu Dobawenata Stoinost). Bulgaria has integrated into its 2006 VAT Act, the VAT rules, Directives,created by the European Union, which Bulgaria joined in January 2007.

[2] This has been adopted in the EU since the removal of fiscal frontiers.

[3] The buffer is a VAT-registered taxable person who is placed in the transaction chain between the missing/defaulting trader and the broker. Depending on the complexity of the fraud, there can be any number of buffers. The broker is a VAT-registered taxable person who sits at the end of the transaction chain and either dispatches or exports the goods or services. As the supply is VAT zero rated the broker incurs an input tax liability but no output tax liability, thus making its VAT return a repayment return from the National Revenue Agency. The term contra trader refers to a VAT-registered taxable person that participates in two separate types of transaction chain during the same VAT period, where the output tax from one chain is designed to off-set the input tax incurred on the other chain.

This report is structured as follows. Section 2 provides a visualization of the VAT network in Bulgaria. Section 3 briefly describes the methodological contribution. Section 4 presents the some of the technical elements of the methodology. Section 5 describes the data, while Section 6 presents the results. Section 7 provides concluding remarks and recommendations.

2 The VAT network of transactions

Traders/taxable persons are interlinked through the production chain, a linkage identified through the invoice-credit mechanism and the paper trail of those transactions discussed above. The total number of registered traders/taxable persons in Bulgaria in the years analysed is 312,762 and, on average, 75% of those traders/taxable persons make at least one transaction in a given month during the time period under investigation.

Figure 1 provides a visualization of the VAT network in Bulgaria, across all of the 19 economic sectors (the list of sectors is provided in Table 1). In the graph, the direction of the arrows (directed edges) depict the direction of the transactions whereas the width of each arrow reflects the total amount of the VAT base (as reflecting in the invoices submitted) between the corresponding pair of sectors.

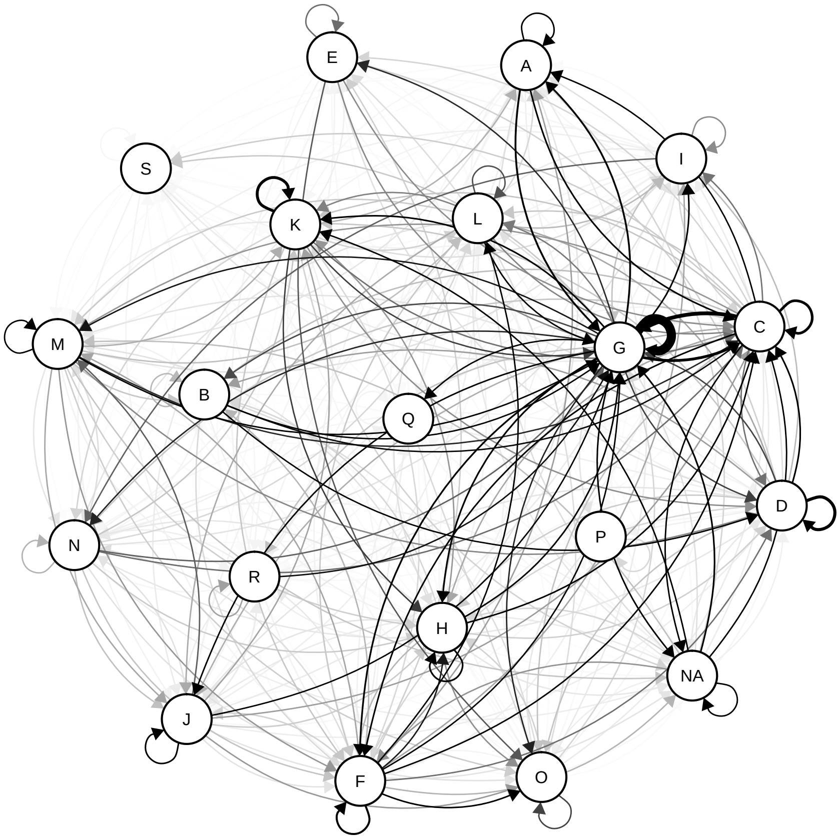

Figure 1: The network of sector-specific transactions in the VAT network. Each node corresponds to an economic sector whereas the edge direction represents sells.

In the complex network of transactions of Figure 1, the objective is to identify groups of traders/taxable persons ('clusters') which conceal transactions and so the amount of VAT due to the National Revenue Agency through fictitious trading and other illegitimate transactions.[4]

3 The methodology explained briefly

The key idea behind the methodology is that the VAT network, and how traders/taxable persons connect through the network (and the intensity of connections), has a 'structure' which when coupled with available information related to identified fraud has the potential of predicting anomalies in the transactions within the structure.[5] The analysis make use of Graph Theory, Machine Learning and Data Analytics.

More specifically (but still briefly), the methodology is described as follows. As noted earlier, VAT is a network where VAT-registered traders/taxable persons are interlinked through economic transactions. All the transactions within the VAT system are represented as edges (connections) in a directed graph in which the vertices correspond to VAT-registered traders. The analysis focuses on the following two types of transaction anomalies (fraud): (i) the detection in which particular traders exhibit transaction deviations from normal patterns,[6] and (ii) detection in which a group of traders creates fictitious transactions to manipulate the financial information submitted to the tax authorities.[7] Identifying both (i) and (ii) above will identify clusters of traders who are engaging in fraudulent transactions.

To perform anomaly detection on the networks the analysis combines recently developed Machine Learning techniques with algorithms that perform clustering detection in networks consisted of a large number of traders. More precisely, first, the likelihood that a VAT-registered traders/taxable person performs fraudulent transactions is estimated.[8] Then, in order to detect small clusters of traders/taxable persons that are involved in illegitimate VAT transactions, the estimated probabilities are used as an input in clustering algorithms that are especially designed to detect densely connected communities in large graphs.[9] This then will be applied on the data provided and which are described in Section 5.

4 Technical details of the methodology

This section presents some of the technical details of the methodology.

4.1 Graph representation

All VAT transactions observed in a given month are represented as a weighted directed graph. A graph is defined as $G = (V, E)$ where V is the set of vertices (nodes) and $E \subset V \times V$ is the set of edges. In a directed graph $G = (V, E)$ every edge $(i, j) \in E$ links node i to node j (ordered pair of nodes). An undirected graph is a directed one where if edge $(i, j) \in E$ then edge $(j, i) \in E$ as well. Every graph $G = (V, E)$ (directed or undirected, weighted or unweighed) can be represented by its adjacency matrix $\boldsymbol{A}$. Matrix $\boldsymbol{A}$ has size $n \times n$,

[4] Transactions might be fictitious, that would be when traders/taxable persons acquire invoices which do not correspond to real transactions (as in the Missing Trader Intercommunity (MTIC) transactions), or they can be incorrectly declared. Both transactions are needed to be identified.

[5] And something that applies more broadly, and it is not specific only to VAT fraud.

[6] This is called *anomalous vertex* detection.

[7] Or completely conceal it. This is called *anomalous sub-graphs.*

[8] Utilizing statistical methods, such as gradient boosting regression (Friedman, 2001).

[9] See for example Sussman et al. (2012) and Binkiewicz et al. (2017) for a detailed discussion on community detection and recent advances.

where n is the number of vertices in the graph, the rows and columns represent the nodes of the graph and the entries indicate the existence of edges. We write

$$\boldsymbol{A}_{ij} = \begin{cases} w_{ij}, & \text{if } (i,j) \in E, \quad \forall\ i,j \in 1,\dots,n \\ 0, & \text{otherwise.} \end{cases}$$

In the case of unweighted graphs w_{ij} is a binary variable indicating the existence of and edge between the ith and jth node and in the case of weighted network w_{ij} is the weight of the edge. If the graph is undirected, the adjacency matrix $\boldsymbol{A}$ is symmetric, i.e., it is equal with its transpose $\boldsymbol{A}^T$, while for directed graphs the adjacency matrix is non-symmetric. See Malliaros and Vazirgiannis (2013) for a detailed presentation of the theory related to graphs, as well as for other notions that include but are not limited to the degree and the strength of the vertices and are commonly used in the analysis of graphs.

4.2 Anomaly detection

4.2.1 Identification of anomalous vertices

To detect the two types of graph-anomalies described in Section 5 we combine recently developed Machine Learning tools with algorithms that perform community detection in Large Graphs. We first focus on identifying anomalous vertices in a given graph and then we utilize this information in order to conduct efficiently detection of anomalous communities.

To identify anomalous vertices we set $\boldsymbol{Y} = (y_1, \dots, y_n)$ to be an n-dimensional binary vector where one indicates that the ith vertex is anomalous, $i = 1, \dots, n$. We also set X_i to be a p-dimensional vectors consisted of the features associated with the ith vertex. For each vertex we utilize two type of features. The first type consists of the features presented in Section 5, and are provided by the National Revenue Agency. The second type includes features from the graphs that we construct from the monthly aggregated VAT base of the declared invoices. These are the strength and the in-and-out degree of each vertex in the month that we aim to detect anomalous vertices as well as their means across the previous months. Under the described set-up the problem of anomalous vertex identification can be seen as a classification exercise. Thus, we perform the desired classification by utilizing gradient boosting regression (Friedman, 2001). In particular, we employ the extreme gradient boosting (XGboost) method developed by Chen and Guestrin (2016). XGboost is an implementation of gradient boosted decision trees designed for speed and performance by utilizing parallel programming. It combines gradient information with boosting techniques to minimize the prediction error in a classification modelling prediction problem. Boosting is an ensemble technique where new models, typically decision trees, are added to correct the errors made by existing models while in gradient boosting a gradient descent algorithm minimizes the loss when new models are added. For a detailed presentation of gradient boosting methods, and XGboost particularly, see for example James et al. (2013). After the application of the XGboost method in our data we obtain the n-dimensional vector $\hat{\boldsymbol{Y}}$ which consists of the predicted vertex specific probabilities of anomalousness.

4.3 Identification of anomalous communities

To detect anomalous communities in a given graph we propose to utilize its structure and weights as well as the available features of each vertex. To achieve this we employ recently developed community detection methods that utilize information about the vertices. We, particularly, explore the method proposed by Binkiewicz et al.

(2017) where the outer product of the matrix with available covariates for the vertices of a undirected and non-weighted network is added to the regularized Laplacian of the graph. In our approach for detection of anomalous sub-graphs we first extend the techniques developed by Binkiewicz et al. (2017) in directed weighted graphs. Then, we utilize the XGboost method as presented in Section 4.2.1 in order to summarize the information from the available vertex specific features in the community detection algorithm. We avoid, thus, computational problems caused by the large number of vertices while we reduce the computational cost of the proposed methodology as well.

Our proposed methodology is summarized as follows. By noting that the adjacency matrix $\boldsymbol{A}$ corresponds to a directed graph we first transform $\boldsymbol{A}$ in a suitable form such that techniques for undirected graph to be applied. Among the various techniques (see for example Malliaros and Vazirgiannis (2013)) we set $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{A}^T$ and we note that the matrix $\tilde{\boldsymbol{A}}$ is symmetric. The corresponding undirected graph has the same edges as the original one but in the case of directed edges in both directions, the weight of the new edge is the sum of the weights of the initial directed edges. Community detection methods that are based on $\tilde{\boldsymbol{A}}$ tend to group nodes that share similar incoming and outgoing edges (Satuluri and Parthasarathy, 2011); this is a feature that we are interested in since it is reasonable to assume that VAT-registered traders that perform fraudulent activity have common trading patterns. Next, following Binkiewicz et al. (2017), we consider the eigendecomposition of the matrix

$$\tilde{\boldsymbol{L}}(\alpha) = \boldsymbol{L}_\tau \boldsymbol{L}_\tau + \alpha \hat{\boldsymbol{Y}} \hat{\boldsymbol{Y}}^T, \tag{1}$$

where $\boldsymbol{L}_\tau$ is the regularized graph Laplacian of the undirected graph with weighted adjacency matrix $\tilde{\boldsymbol{A}}$ and τ is a constant that improves spectral clustering performance on sparse graphs; see Binkiewicz et al. (2017) for details. Finally, α is a positive tuning parameter chosen to achieve a balance between $\boldsymbol{L}_\tau$ and $\hat{\boldsymbol{Y}}$ such that the information in both is captured in the leading eigenspace of $\tilde{\boldsymbol{L}}(\alpha)$. From the eigendecomposition of $\tilde{\boldsymbol{L}}(\alpha)$ we obtain the K eigenvectors that correspond to K largest eigenvalues of the matrix. Finally, we assign each vertex to one of K clusters by treating each normalized eigenvector as a point in $\mathbb{R}^K$ and running the k-means algorithm with K clusters. Algorithm 1 summarizes the steps of our proposed method.

5 The data

The administrative data are VAT returns and ledgers which cover the universe of VAT transactions of Bulgarian traders/taxable persons. The data consists of all records required traders/taxable persons to declare under the Bulgarian VAT law (for example, Domestic Transactions/Imports/Exports/Inter-community Acquisitions (ICA)/Inter-community Deliveries (ICD)/Special Aquisions/Reduced rates/Triangural Acquisitions (TA)/Triangular Deliveries (TD)). The ledgers contain the unique identifier of the sellers/buyers and the value of each transaction (invoice). The data required significant pre-processing in order to be cleaned and be ready for statistical analysis. Given the time constraint imposed on us by the project, the methodology is applied to the years 2016 and 2017.

The constructed graphs are based on the aggregation of the VAT base from all deliveries between each pair of VAT-registered traders/taxable persons. The total number of VAT-registered traders/taxable persons in Bulgaria is 312, 762 and, on average, 75% of them conduct at least one transaction in a given month. Table 1 reports the VAT base in each of the two years, as well as the composition of VAT base according to the categories.[10]

Table 1: The total VAT base reported on sells invoices and imports for the years 2016 and 2017 across the categories of VAT transactions.

	2016	2017
Sum of VAT base (sells and 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 out of Bulgarian territory (%)	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

Table 2 provides the summary statistics of the categories of transactions over the 24-month period.

Table 2: Total number of transactions.

	Median	Minimum	Maximum	Standard deviation
ICA	390,843	147,787	528,678	107,988
ICD	217,397	85,782	305,091	60,130
9%	193,019	91,969	251,009	39,338
Services from EU	212,614	64,362	284,953	56,329
Deliveries from out of Bulgarian territory	84,360	18,529	168,261	46,533
Exports to third countries	280,437	96,196	413,784	88,778
Imports from third countries	49,237	14,797	65,271	13,308
0% special deliveries	8,148	2,801	14,733	3,237
TA	12,480	2,619	16,244	4,250
TD	13,023	1,689	18,067	4,365

Table 1 reports the percentages for the categories of VAT transactions. Clearly, for the years 2016/2017, ICA and ICD dominate VAT transactions in Bulgaria. Figure 2 displays the sum of VAT base reported on the sells invoices at each month of the years 2016 and 2017.

[10] All values are expressed in local currency.

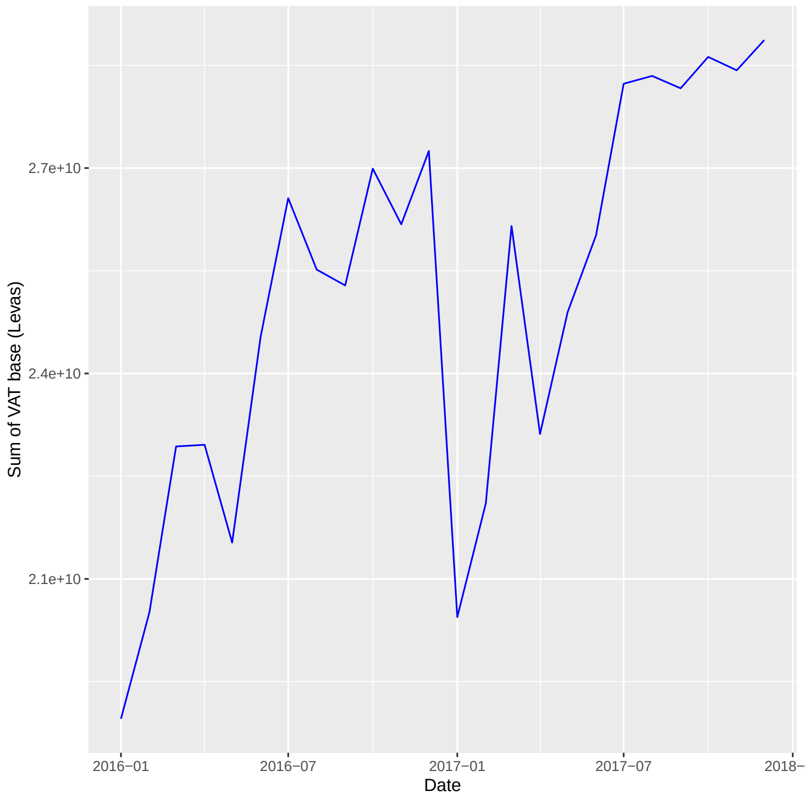

Figure 2: Total VAT base reported on the sells invoices at each month of the years 2016 and 2017.

In addition to the aggregated VAT base of the invoices, the analysis utilises a set of features that describe the profile of each VAT-registered trader/taxable person. 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/taxable persons.[11] Importantly, each registered taxpayer has been classified as high risk or low risk based on criteria developed by the Revenue Agency which utilises operational knowledge and past information of fraudulent activity. 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/taxable persons is also available and classified according to the categories displayed in Tables 1 and 2.

Figure 3 presents the distribution of the VAT-registered traders/taxable persons across the economic sectors in years 2016 and 2017[12]

[11] The Appendix provides details and descriptive statistics for all the additional variables related to the profile of the VAT-registered traders/taxable persons.

[12] The economic sectors are provided in Table 6. Figure 3 provides the distribution of VAT-registered traders/taxable persons across all sectors. As can be seen from this figure Sector G (Wholesale and Retail Trade) is the largest sector (excluding sector NA (Not Available) which contain all those traders/taxable persons who have not declared an economic activity sector).

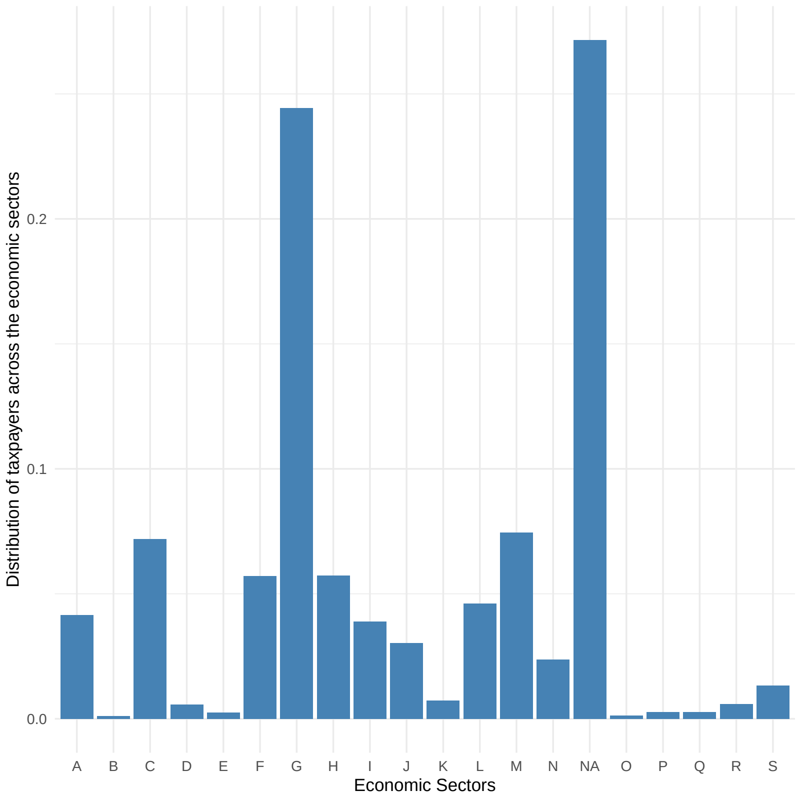

Figure 3: Distribution of VAT-registered traders/taxable persons across economic sectors (of Table 6).

But specific types of transactions differ across sectors. Figure 4 presents this.

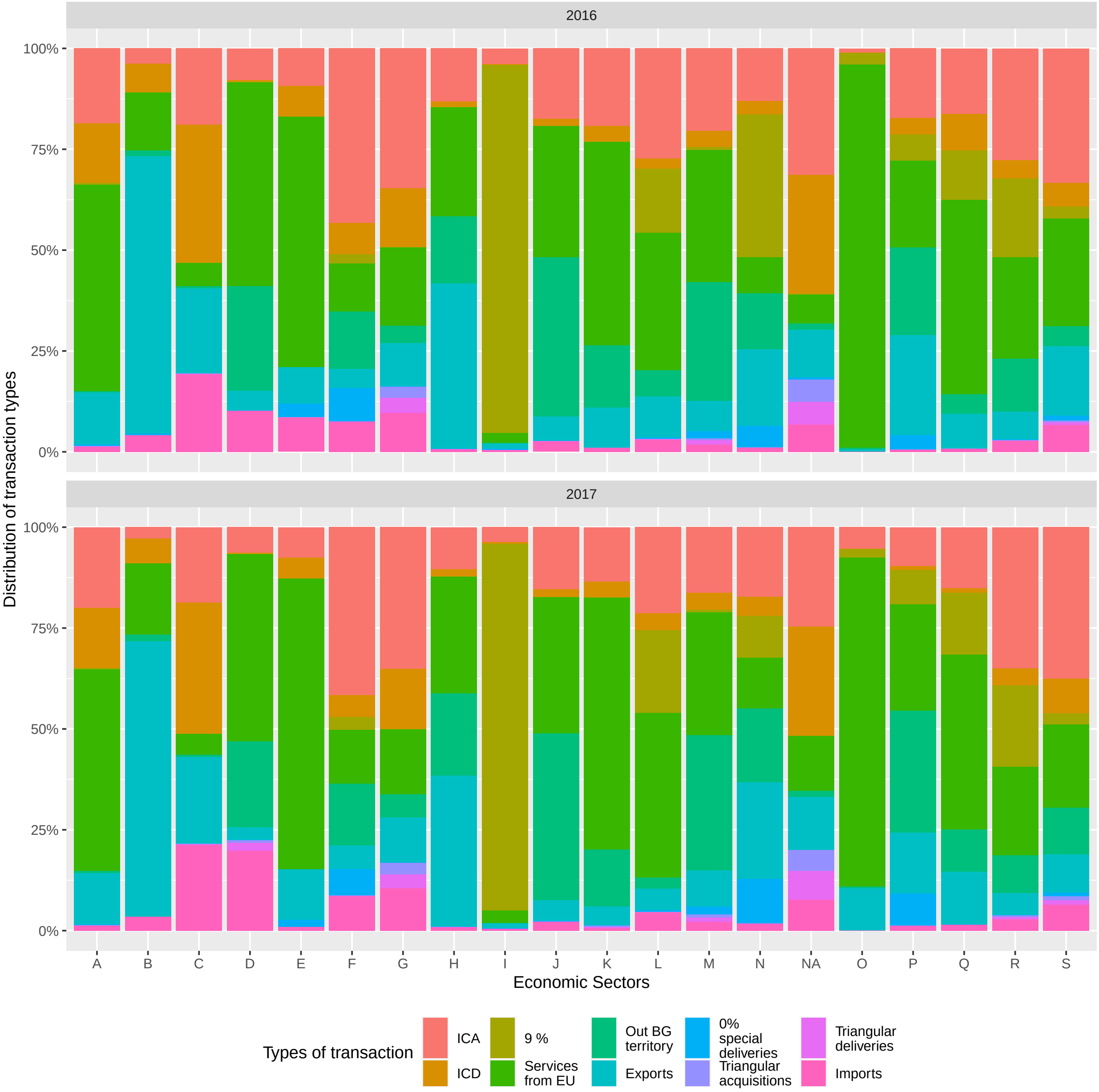

Figure 4: Distribution of types of transactions within each economic sector (NACE Classification Codes appear in Table 6).

As noted, the National Revenue Agency has provided information on the high risk sectors (categorised as in Table 6). Figure 5 presents the empirical risk probabilities within each economic sector.

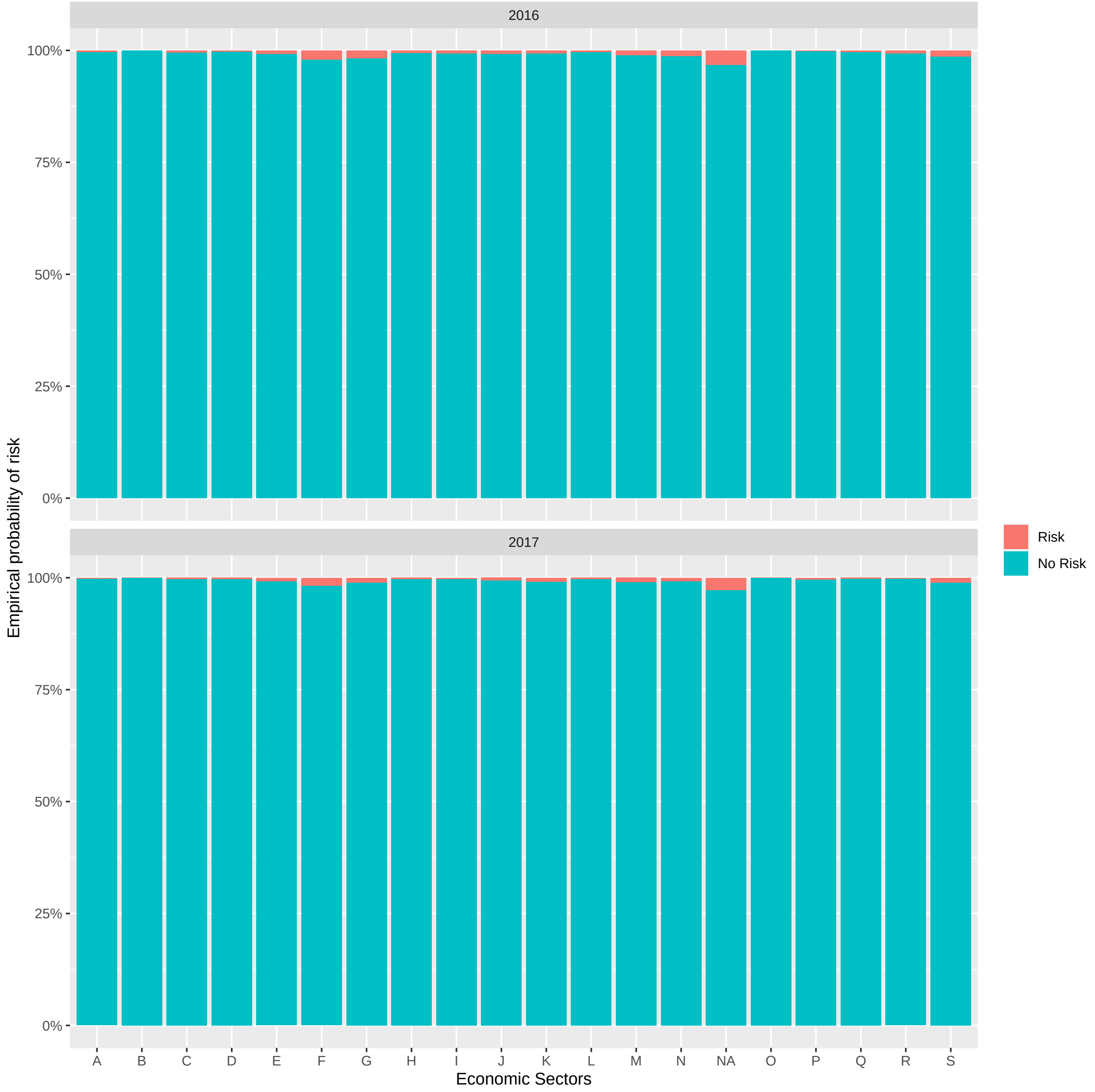

Figure 5: Empirical risk probabilities within each economic sector (of Table 6).

Our proposed methodology is based on the construction of graphs that represent the VAT transactions of each month. For a given month we utilize data from all invoices issued by all VAT-registered traders/taxable persons in Bulgaria to other VAT-registered Bulgarian traders/taxable persons in order to construct the graph of the observed transactions.

The next section discusses the results.

6 Results

The methodology developed is applied on the data described in Section 5 for the period 2016 and 2017. We have constructed 24 graphs, each one with $n = 312,762$ number of VAT-registered Bulgarian traders, similar to Figure 1, in order to represent the observed monthly VAT transactions of the years 2016 and 2017.

To evaluate the proposed methodology we conducted the following exercise. Utilising the features of the VAT-registered Bulgarian traders, as well as the given classification of high risk and low risk traders up to November 2017 provided by the National Revenue Agency, we identify traders/taxable persons and clusters which have engaged in VAT fraudulent transactions in December 2017. The number of VAT-registered traders/taxable persons during that period is $207,412$. Of note is that the identified fraudulent traders for December 2017 were $2,192$ but due to limitations of the auditing processes performed by the National Revenue Agency it is possible that existing traders/taxable persons mights have been misspecified as non-fraudulent. Thus, by noting that the proposed community detection methods are guided by the probability of a trader/taxable person to be fraudulent, we are able to report traders that have been not identified by the National Revenue Agency as such but could be further investigated being involved in fraudulent activities. Table 7 displays summary statistics for the 'degrees' and the 'strengths' of the network constructed by considering an edge between two (nodes) traders/taxable persons if there is at least one sell invoice exchanged.[13] The weight of the edge is the VAT base reported on the invoice. Table 8 shows the corresponding statistics for the traders/taxable persons already identified as high risk by the National Revenue Agency.

Application of the method for anomaly detection developed identifies 191 clusters which consist of more than one VAT-registered traders/taxable person. In particular, 70% of the identified clusters have 10 or less members of VAT-registered traders/taxable persons, 25% of the clusters have size between 10 and 100 while there are 5 clusters with more than 100 members but less than $1,000$ and 2 clusters with size greater than $1,000$ of VAT-registered traders/taxable persons. The largest of the identified cluster contains 94% of the VAT-registered traders/taxable persons that were active in December 2017. Interestingly, this cluster includes only 200 out of the $2,192$ traders/taxable persons that have been marked as high risk by the National Revenue Agency. Thus, we consider this as the cluster with legitimate traders/taxable persons, taking that the rate of false negatives of the proposed methodology is slightly less than 10%.

Excluding this big cluster the remaining 190 clusters have in total $10,624$ of VAT-registered traders/taxable persons. Since our method has been trained to identify clusters with high risk VAT-registered traders/taxable persons these clusters are signified as groups where fraudulent activity occurs. Importantly $2,016$ of the $10,624$ traders/taxable persons included in the 190 clusters have already been identified as high risk from the National Revenue Agency which implies that true positive rate of our method is $2016/2192 = 92\%$. It is worth emphasising that there are more than $8,000$ VAT-registered traders/taxable persons in the clusters that have not been identified as being involved in fraudulent transactions. In Figure 6 we display the proportion of the known to the National Revenue Agency high risk traders/taxable persons included in the identified clusters that contain at least one non-legitimate taxpayer.

[13] The degree is an indication of how many nodes a VAT-registered trader/taxable person within a cluster is connected to, whereas the strength signifies how much these trader/taxable persons transact. We return to this later on.

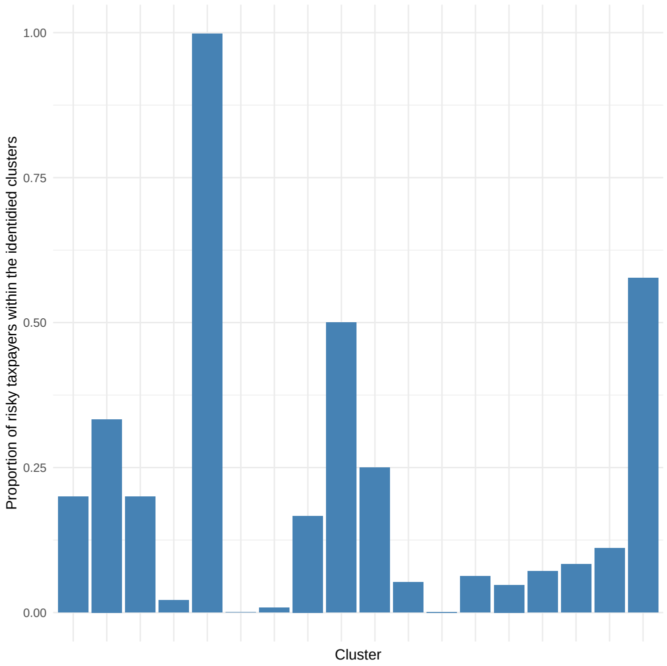

Figure 6: Proportion of traders/taxable persons that are already identified by the tax authorities as non-legitimate within each cluster. We display the proportions for the 18 clusters which include at least one non-legitimate taxpayer.

The Tables 3, 4 and 5 as well as the Figure 7 present summary statistics for the degrees and the strengths of the nodes within the identified clusters. More precisely, Table 3 displays statistics for the biggest identified cluster, with $196,237$ members, which, as explained above, is considered as the cluster with legitimate traders/taxable persons.

Table 3: Summary of characteristics of nodes within the cluster with traders/taxable persons considered as legitimate.

	1st quartile	median	3rd quartile
Degree	3.00	10.00	24.00
Strength	1,399.10	10,277.70	40,378.49

Tables 4 and 5 and Figure 7 show the corresponding statistics for the rest 190 clusters which are assumed to contain high risk VAT-registered traders/taxable persons. More precisely in Table 4 we display the characteristics of the nodes in two relatively large clusters, with 397 and $1,752$ members, in which the proportion of known non-legitimate VAT-registered traders/taxable persons is high. Table 5 presents node characteristics for 13 smaller clusters, less than 100 members and include at least one known non-legitimate VAT-registered trader/taxable person. Figure 7 shows box plots for the rest 170 clusters in which all the traders/taxable persons have been not identified as non-legitimate from the National Revenue Agency.

Table 4: Summary of characteristics of nodes within the 2 identified clusters that contain the majority of the non-legitimate VAT-registered traders/taxable persons. The first row in degrees and strengths corresponds to a cluster with 397 members in which 60% of them are non-legitimate and the second row corresponds to a cluster with $1,752$ members in which 99% of them are non-legitimate.

	1st quartile	median	3rd quartile
Degree	2.00	7.00	14.00
	3.00	8.50	22.00
Strength	601.37	6,911.75	30,137.55
	1,201.12	9,489.28	46,035.41

As noted earlier, the degree is an indication of how many nodes a VAT-registered trader/taxable person within a cluster is connected to, whereas the strength signifies how much these traders transact. Take, for example, the last line in the 2 blocks of Table 5 which shows that in the 13-th cluster there are between 15 and 30 traders which are connected through the invoice-credit mechanism. In terms of VAT amounts this ranges from 992,412 to 1,433,162.44.

Table 5: Summary of characteristics of nodes within the 13 identified clusters with less than 100 and include at least one non-legitimate registered taxpayer.

	1st quartile	median	3rd quartile
Degree	36.00	48.00	106.00
	22.00	32.00	175.50
	2.00	3.50	12.50
	29.00	41.00	82.00
	22.00	54.00	81.75
	64.50	174.50	600.75
	57.00	58.00	113.00
	61.50	122.00	197.50
	9.00	138.00	223.00
	18.00	40.00	77.00
	3.00	5.00	7.75
	3.50	6.00	8.50
	15.00	22.00	30.00
Strength	3,693,972.92	4,725,788.98	28,607,114.73
	735,020.23	1,239,706.58	1,910,732.21
	639,539.33	999,406.78	1,668,842.79
	300,652.26	384,944.20	496,689.05
	656,656.65	1,167,968.00	2,406,773.92
	3,336,130.48	5,335,173.97	10,913,038.80
	7,863,060.76	8,948,021.82	11,403,730.16
	2,870,157.47	3,934,883.61	4,948,072.73
	7,014,573.40	8,177,396.82	10,881,038.55
	1,007,572.40	1,791,405.99	3,391,747.04
	1,126,956.60	1,156,010.02	1,797,316.93
	22,797,799.01	2,2810,768.82	22,823,738.62
	992,412.67	1,213,910.29	1,433,163.44

Figure 7 presents information on the distribution of degree and strength. As can be seen from this figure, the degree varies less across the quartiles than the strength. This is important information that can be utilised in understanding the characteristics of the VAT network.

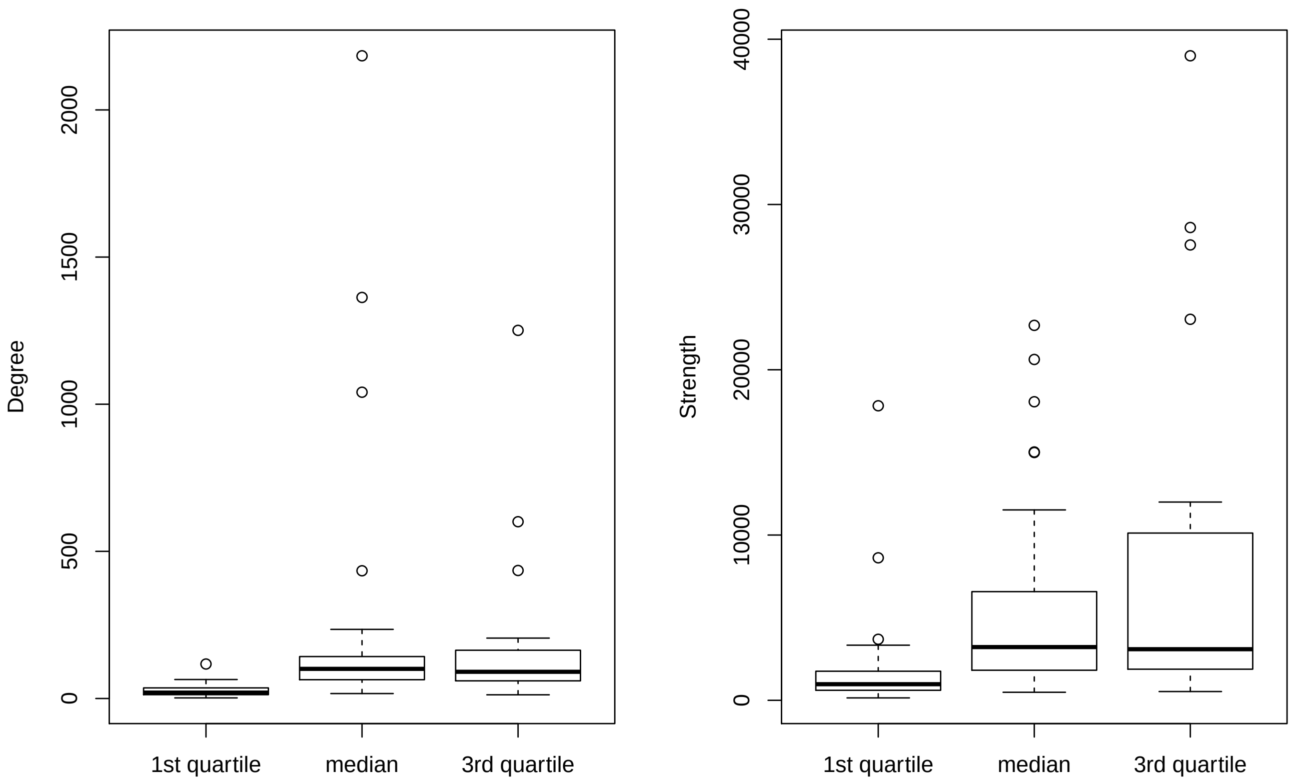

Figure 7: Box plots with summary statistics for the degrees (left) and strengths (right) of the nodes in clusters that do not include any known non-legitimate traders/taxable persons.

Finally, to further evaluate the method and approach a random sample of 35 out of the 8, 000 traders/taxable persons which have not been classified as high risk by the National Revenue Agency were given to the National Revenue Agency for investigation and evaluation. The National Revenue Agency confirmed that out of those 35 VAT-registered traders/taxable persons 12 were identified as high risk, without however placing a number on revenues forgone.

7 Concluding remarks and recommendations

A powerful method has been developed which has the potential of combating VAT fraud and closing the VAT Gap in Bulgaria. The analysis has identified more than 90% of the high risk VAT-registered traders/taxable persons the National Revenue Agency has identified as such. It has also identified 12 high risk VAT-registered traders/taxable persons out of a sample of 35 drawn from the 8000 not identified as high risk by the National Revenue Agency. The nature of VAT fraud is dynamic in nature and the method is flexible enough to capture changes in the type and size of the fraud. There are a number of recommendation which are coming out of this analysis:[14]

[14] Implicit in the analysis is the issue of revenue yield, from the identified fraudulent clusters. This is required to be made more specific and further analysis to explore this is needed.

- The analysis in this report must be repeated using more recent VAT transaction data (and also over more years), so the methodology can be fine-tuned and further tested on the available data.
- A prerequisite for applying the methodology is that the risk scoring is sufficiently and accurately updated. There is an inherent feedback from the risk scoring to the identification of fraudulent transactions, through the evaluation of predicted fraudulent clusters.

Appendices

Appendix A

This appendix presents some additional graphs.

Table 6: NACE Classification Codes.

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 bodies

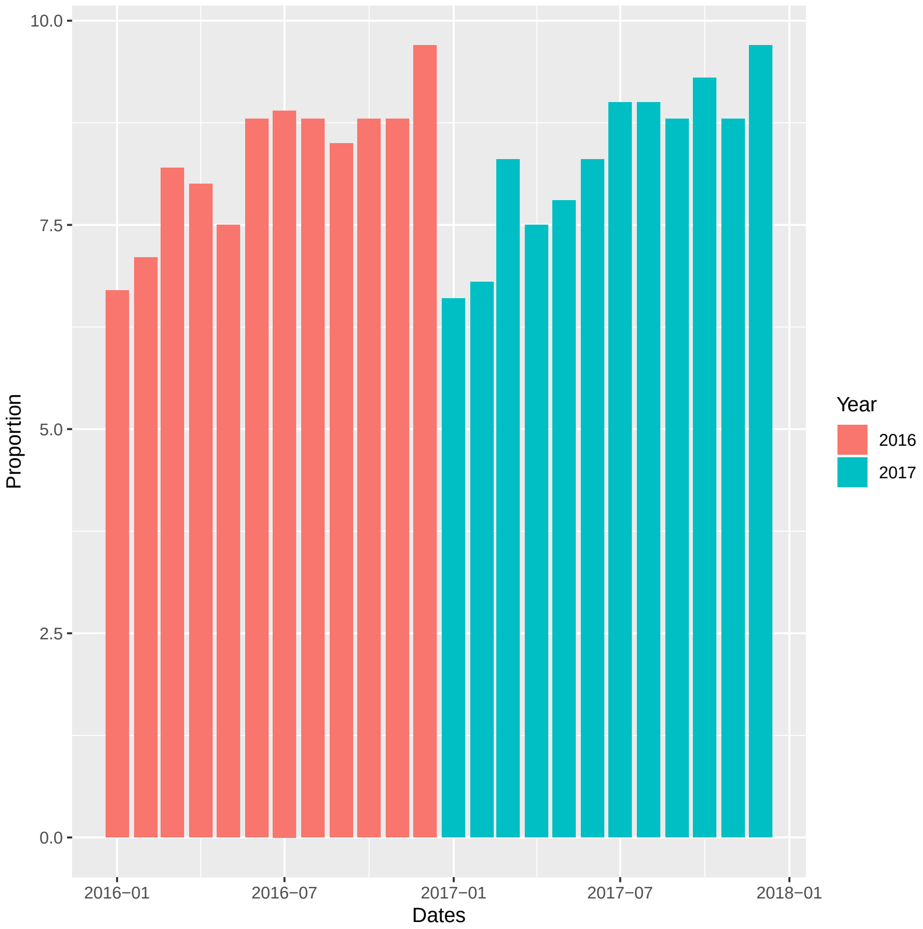

Figure 8: Monthly proportions of the VAT base reported on the sells invoices of the years 2016 and 2017.

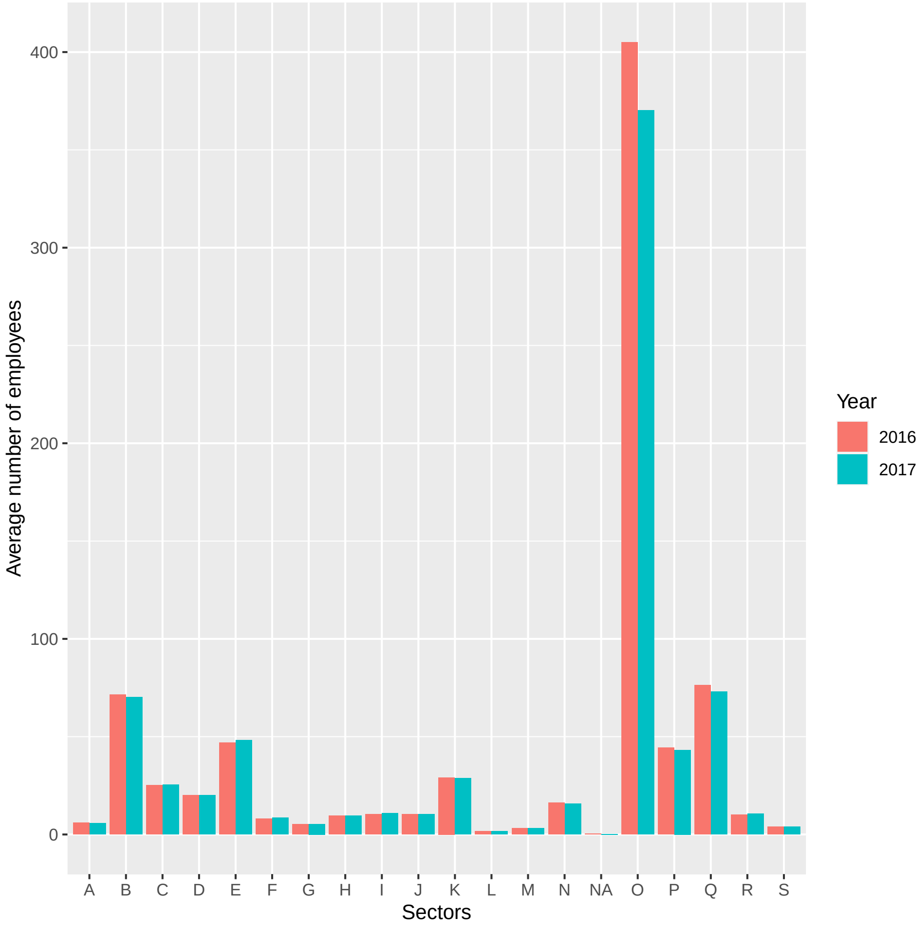

Figure 9: Average number of employees across the companies of each sector; see Table 6 for the NACE Classification Codes.

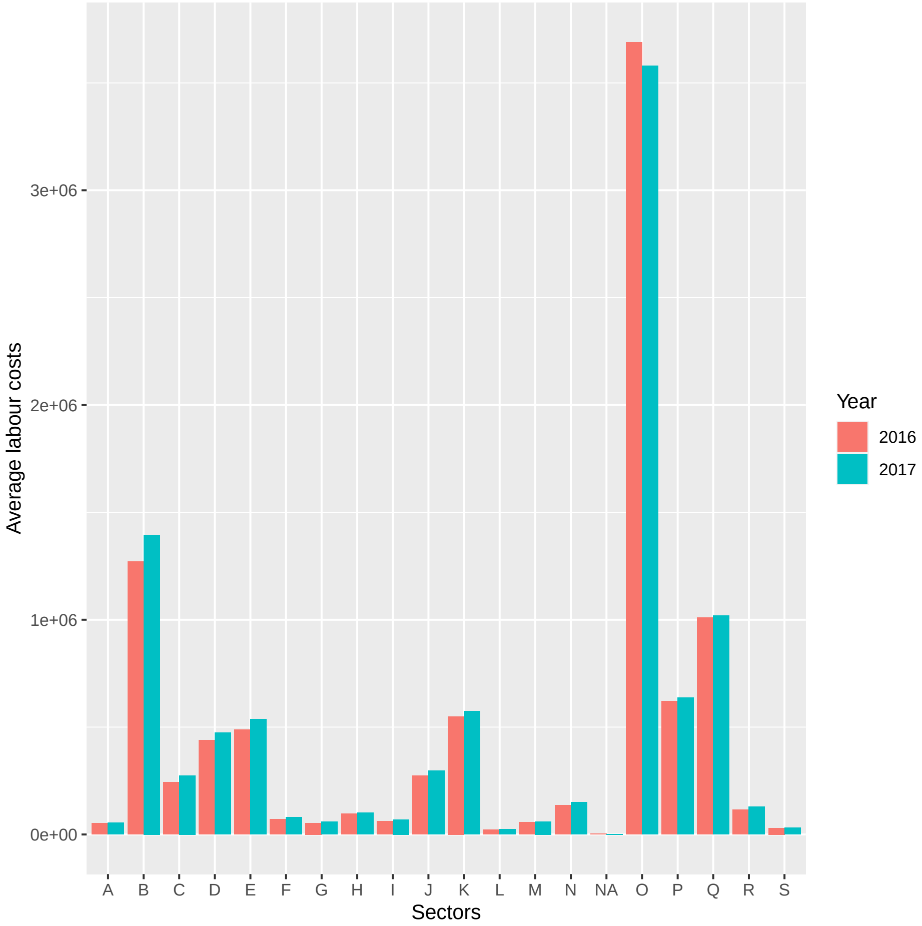

Figure 10: Average number of labour costs across the companies of each sector; see Table 6 for the NACE Classification Codes.

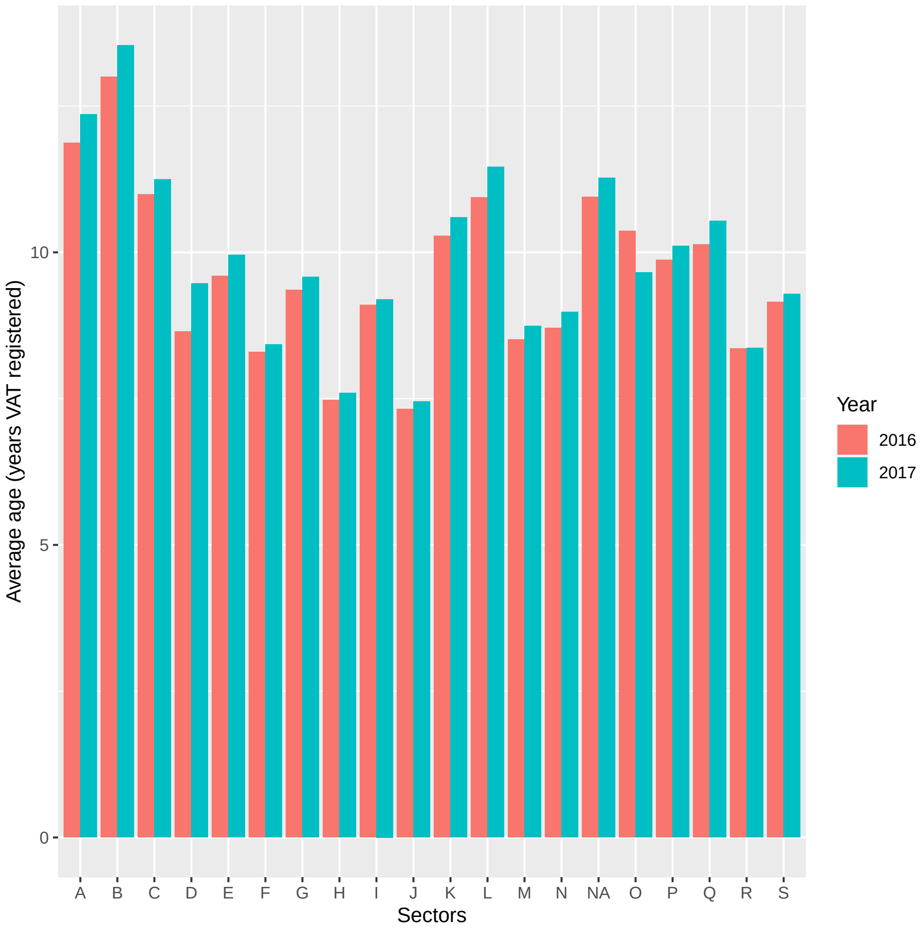

Figure 11: Average number of years of registration across companies of each sector; see Table 6 for the NACE Classification Codes.

Table 7: Summary of node characteristics

		1st quartile	median	3rd quartile
Degree	All	3.00	10.39	25.91
	In	2.70	7.87	18.48
	Out	0.00	1.00	4.96
Strength	All	1,427.23	10,605.48	44,317.01
	In	689.89	5,381.59	25,733.97
	Out	0.00	1,008.66	11,318.33

Table 8: Summary of node characteristics for traders/taxable persons marked as high risk.

		1st quartile	median	3rd quartile
Degree	All	2.00	3.74	8.20
	In	1.00	2.00	4.04
	Out	0.00	1.00	3.13
Strength	All	1,080.15	31,924.19	153,518.33
	In	113.72	2,167.26	31,532.27
	Out	0.00	14,431.89	94,929.87

Appendix B

Appendix B presents the algorithm.

Algorithm 1 Anomalous community detection

```
Input: $n$-dimensional vector $\hat{\boldsymbol{Y}}$ with predicted probabilities of anomalousness; the regularized graph Laplacian $\boldsymbol{L}_\tau$ that corresponds to the transformed weighted adjacency matrix $\tilde{\boldsymbol{A}}$; number of clusters $K$; tuning parameter $\alpha$.
: Set $\tilde{\boldsymbol{L}}(\alpha) = \boldsymbol{L}_\tau \boldsymbol{L}_\tau + \alpha \boldsymbol{Z}\boldsymbol{Z}^T$
: Compute the eigendecomposition $\tilde{\boldsymbol{L}}(\alpha)$.
: Form the $n \times K$ matrix $\boldsymbol{U}$ with columns the eigenvectors that correspond to the $K$ largest eigenvalues.
: Normalize each row in $\boldsymbol{U}$ to have unit length.
: 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 $i$th row of $\boldsymbol{U}$ falls in the $k$th cluster assign node $i$ to cluster $k$.
```

Appendix C

The routines of the methodology are available upon request.

References

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

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.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

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

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

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.

Sussman, D. L., M. Tang, D. E. Fishkind, and C. E. Priebe (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association 107*(499), 1119–1128.