Detecting Network Anomalies in the Value Added Taxes (VAT) system

Angelos Alexopoulos[1] Petros Dellaportas[2] Stanley Gyoshev[3] Sofia Olhede[4]

Christos Kotsogiannis[1,5] Trifon Pavkov[6]

[1]University of Exeter and TARC, UK
[2]UCL and Alan Turing Institute, UK and AUEB, Greece
[3]University of Exeter, UK
[4]EFPL, Switzerland
[5]CESIfo, Germany
[6]Bulgarian National Revenue Agency

June 24, 2020

The views are those of the authors and should not be attributed to the Bulgarian National Revenue Agency (NRA). The research has been subject to a confidentiality agreement between the Researchers and the NRA and no taxpayer individual information has been disclosed to the Researchers. Financial support from HSBC-Alan Turing Institute under TEDSA2/100056 is gratefully acknowledged

Road map

- Motivation the topic and research questions
- A bit on VAT
- Description of data used in the analysis
- Description of methodology
- Results and evaluation
- Conclusion

Motivation

- Research is motivated by the significant ‘fraud’ in Value Added Tax (VAT)
- Difficult to obtain accurate estimates—some have it that VAT fraud in EU is around 50 billion Euros (lower bound)
- Revenue Authorities do utilise algorithms, but there is scope for academic work and cooperation with such organisations
- Objective of research:
    - Develop a model which is fed with information (‘and trained’) to predict ‘high risk’ behaviour but also identify the cluster this ‘high risk’ behaviour belongs to (sub-network/cluster)
    - The model is applied to VAT but idea is more broadly applicable

VAT: Main elements

- VAT is a **broad-based tax on consumption** and has dominated the world (as considered to be an 'efficient' tax system)
- Explicit **credit-invoice** mechanism where firms/taxable persons
    - **Levy** VAT on their output
    - **Deduct** VAT already paid on inputs, and
    - **Remit** the balance due to the government
- In one level, credit-invoice mechanism **facilitates enforcement** as it creates a **paper-trail** of transactions. . . but. . .
- Being a consumption tax, **exports are not taxable** and tax payments are subject to periodic declaration by firms
    - And this is the **Achilles' heel** of VAT—which is duly exploited by unscrupulous traders
- VAT fraud is complicated, sometimes involving dozens of firms spanning across countries/continents

- There are so opportunities for fraud. . . for example the Missing Trader. . .

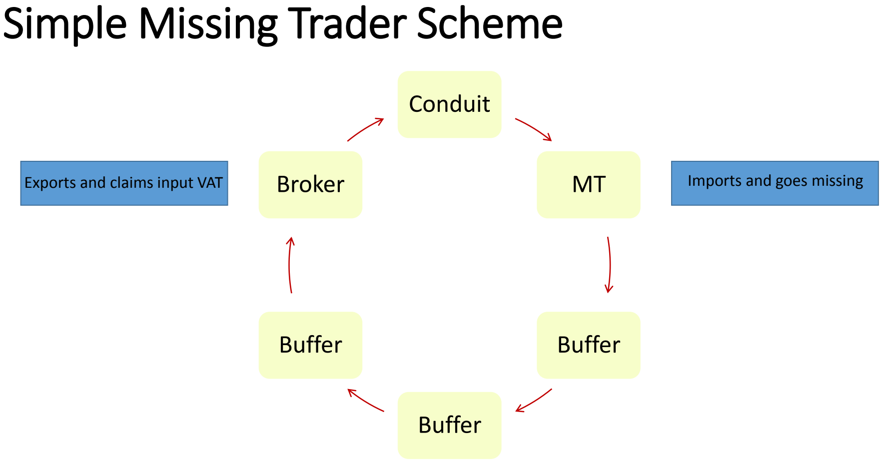


Figure 1: Missing Trade/Carousel fraud

Revenue Authority 'sees' this Network

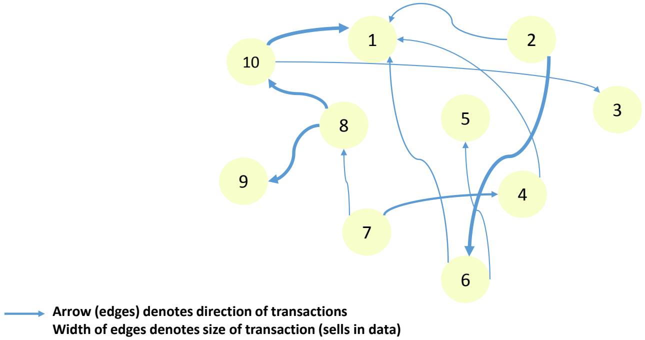


Figure 2: What Revenue Agency 'Sees'

But Real Network is this...

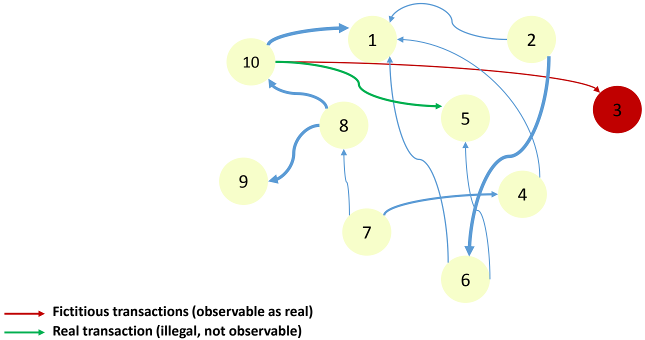
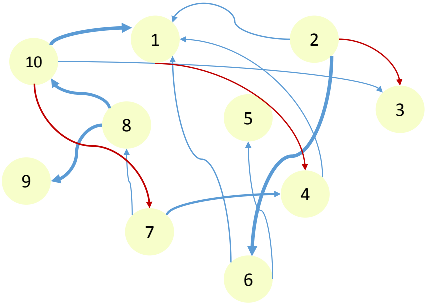


Figure 3: What Revenue Agency does not 'See'

But it might be this the case too!



Fictitious transactions (not real but paper transaction—trading in 'invoices')

Figure 4: What Revenue Agency does not 'See'

- **Recover losses**...difficult [once fraud is done... it is done]
- **Disrupt** the fraud before it begins!
  - This is where we come in... through... trying to **identify** whether there are
    1. Particular taxpayers (vertices) **evolve irregularly** compared to the other vertices ('anomalous vertex detection'), and/or
    2. Groups of taxpayers (vertices) with transactions that **deviate from normal patterns** ('anomalous sub-graphs detection')

Data set: VAT network of transactions in Bulgaria

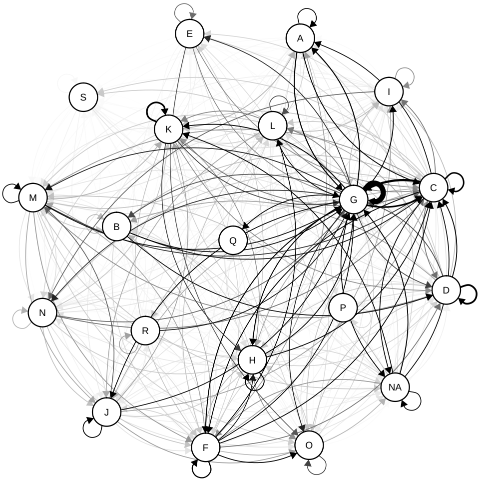


Figure 5: Sector-specific transactions: nodes correspond to economic sectors; edge direction represents sells

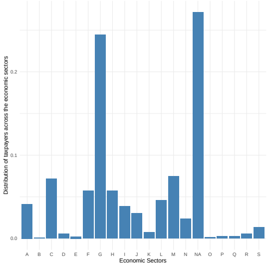


Figure 6: Distribution of VAT-registered traders/taxable persons across economic sectors.

Description of data

- Access to the world of VAT transactions in Bulgaria involving
  - Domestic Transactions/Imports/Exports
  - Inter-community Acquisitions (I.C.A) and Deliveries (I.C.D)
  - Special acquisitions at reduced rates
  - Triangular Acquisitions (TA) and Deliveries (TD)
- VAT returns for all the monthly observed VAT transactions
  - $N = 312,762$ registered taxpayers; 75% active each month
  - 1% of taxpayers are classified as **highly risky** (criteria developed by operational knowledge at NRA and past information)
  - Average monthly transactions: $1,461,198$
- Access to firm specific data: size, age of business, labour costs, sector it belongs to and the...
- Empirical probability of **risk identified** by NRA of firms in a sector

- Monthly VAT transactions are modelled as a **weighted directed graph** where
  - Each vertex/node corresponds to a VAT registered taxpayer
  - An edge between two taxpayers exists if they have exchanged at least one invoice (the direction of the edges represents sells)
  - Edge weights: The sum of the VAT base in all the sells invoices exchanged between two taxpayers
- Network notation:
  - A graph is defined as $G = (V, E)$: V is the set of vertices (nodes) and $E \subset V \times V$ is the set of edges
  - **A** denotes the $N \times N$ adjacency matrix of the graph

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

- **Y** denotes an N-dimensional binary vector that indicates risky taxpayers
- **Aim:** Given the monthly observed VAT networks and the vector **Y** we want to identify individuals and groups taxpayers that perform fraudulent activity in the current month
- We work with data from January 2016 to November 2017 and we **test** the methods in detecting the fraudulent activity in December 2017

Fraud detection

- **Proposed approach:** Utilize the available **node-specific information** (taxpayer profile) to identify **high risk taxpayers** as well as **communities of taxpayers** involved in fraudulent activities
- We develop a two-step method:
  1. We use binary logistic regression to predict risk probabilities for each node
  2. We employ the predicted risk probabilities to perform community detection

Stage 1: Fraud Detection (identifying 'anomalous' nodes)

- We construct the $N \times p$ matrix $\mathbf{X}$ with p node-specific characteristics
- For the ith taxpayer the ith row X_i consists of:
  - Number of transactions and the corresponding VAT base within categories in Tables 1 and 2: ICA, ICD, 9%, Imports/Exports...
  - Company's size, age, time of VAT registration, labour costs, sector
  - **Number of transactions and the corresponding VAT base with highly risky taxpayers**
  - **Averages across months of the graph characteristics: in- and out-degree, in- and out- strength and centrality measures**
- We consider the data set $\{\mathbf{X}, \mathbf{Y}\}$ to train a binary regression model by using extreme gradient boosting regression (XGboost, Chen and Guestrin, 2016)
- We use the trained regression model to obtain the N-dimensional vector $\hat{\mathbf{Y}}$ which consists of predicted node-specific risk probabilities

- We conduct **community detection** taking into account the probabilities $\hat{\mathbf{Y}}$
- We follow (Bienkiewicz et al., 2017) and we perform **spectral clustering** on the matrix
$$\tilde{\mathbf{L}}(\alpha) = \mathbf{L}_\tau \mathbf{L}_\tau + \alpha \hat{\mathbf{Y}} \hat{\mathbf{Y}}^T,$$
where
$$\mathbf{L}_\tau = \mathbf{D}_\tau^{-1/2} \tilde{\mathbf{A}} \mathbf{D}_\tau^{-1/2}, \quad \mathbf{D}_\tau = \mathbf{D} + \tau \mathbf{I}_N$$
- $\mathbf{D}$ is $N \times N$ diagonal matrix where $\mathbf{D}_{ii} = \sum_{j=1}^{N} \tilde{\mathbf{A}}_{ij}$
- $\tau = \frac{1}{N} \sum_{i=1}^{N} \mathbf{D}_{ii}$ is the average node degree and accounts for large nodes and sparse graphs
- $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{A}^T$:
  - $\tilde{\mathbf{A}}$ is symmetric and is the adjacency of the corresponding undirected graph
  - We keep the same edges with $\mathbf{A}$
  - Edges in both directions replaced with one weighted by their sum
- $\alpha > 0$: tuning parameter compromising between the network structure and the probability of fraud

Inputs Graph G with N nodes, $N \times p$ matrix X with node-specific characteristics, spectral tuning parameter $\alpha > 0$

1. Run the XGboost algorithm to obtain node-specific risk probabilities $\hat{\mathbf{Y}}$
2. Construct the matrix $\tilde{\mathbf{L}}(\alpha)$
3. Compute the eigendecomposition of $\tilde{\mathbf{L}}(\alpha)$
4. Form the $N \times K$ matrix $\mathbf{U}$ with columns the eigenvectors of the K largest eigenvalues
5. Normalize each row in $\mathbf{U}$ to have unit length
6. Treat each normalized row of $\mathbf{U}$ as point in $\mathbb{R}^K$ and run a k-means clustering algorithm with K clusters
7. If the ith row of $\mathbf{U}$ falls in the kth cluster assign node i to cluster k

Outputs K clusters which include the nodes of the graph G, node-specific risk probabilities

- We identified $K = 191$ clusters with at least two members in each one
  - 70% of the identified clusters had 10 or less members
  - 25% of the clusters have size between 10 and 100
  - 5 clusters with more than 100 members but less than $1,000$
- The largest cluster contains 94% of the VAT registered taxpayers:
  - Includes only 200 out of 2,192 taxpayers marked as high risk by the authorities
  - We consider this as the cluster with legitimate taxpayers
  - This implies less than 10% rate of false negatives
- The remaining 190 clusters have in total $10,624$ taxpayers; $2,016$ of them already identified from the authorities implying 92% true positive rate of our method

Results: Identifying known fraudsters and testing the method

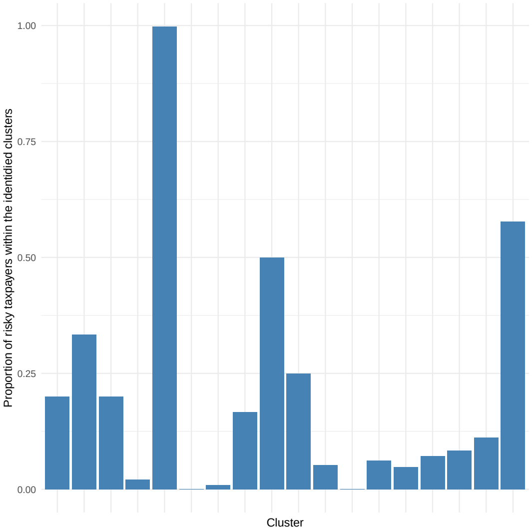


Figure 7: Proportion of VAT registered taxpayers persons that are already identified by the tax authorities as non-legitimate within each cluster. We display the proportions for the 18 clusters which include at least one non-legitimate taxpayer.

- $8,608$ taxpayers in the 190 clusters have not been identified as non-legitimate from the authorities
- We choose 35 (practical restrictions) to be further investigated from the authorities as follows:
  - We rank the $8,608$ taxpayers by using the **predicted node-specific risk probabilities** and we select the first 10
  - To select 15 more we **rank the clusters** that contain at least one known fraudster **by using the mean risk probability within each cluster**; we choose the 15 first clusters and from each one we select the taxpayer with the highest risk probability
  - We select the last 10 by following the same procedure for clusters but consisted completely of **unknown fraudsters**
- Tax authority has reported that 12 out of the 35 VAT-registered traders/taxable considered as high risk (but not £value has been given)

Conclusions

- VAT fraud is significant
- Project has developed a method that identifies clusters of fraudulent transactions
- Limitation: Characterisation of size of fraud across clusters is needed as Revenue Authorities are capacity constrained (we are working on this)

Finally

Thank you for listening!

Please send questions to c.kotsogiannis@exeter.ac.uk