Big Data Analytics for Nabbing Fraudulent Transactions in Taxation System

Priya Mehta[1], Jithin Mathews[1], Sandeep Kumar[3], K. Suryamukhi[1], Ch. Sobhan Babu[1(✉)], and S. V. Kasi Visweswara Rao[2]

[1] Indian Institute of Technology Hyderabad, Telangana, India
{cs15resch11007,cs15resch11004,cs17mtech01002,sobhan}@iith.ac.in
[2] Department of Commercial Taxes, Government of Telangana, Telangana, India
svkasivrao@gmail.com
[3] Plianto Technologies, Telangana, India
cs15mtech11017@iith.ac.in

Abstract. This paper explains an application of big data analytics to detect illegitimate transactions performed by fraudulent communities of people who are engaged in a notorious tax evasion practice called *circular trading*. We designed and implemented this technique for the commercial taxes department, government of Telangana, India. This problem is solved in two steps. In step one, the problem is formulated as detecting fraudulent communities in a social network, where the vertices correspond to dealers and edges correspond to sales transactions. In step two, specific type of cycles are removed from each fraudulent community, which were identified in step one, to detect the illegitimate transactions. We used *RHadoop* framework for implementing this technique.

Keywords: Data mining · Social network analysis · Big data · Goods and Services Tax · Fraud detection · Circular trading · Fraudulent transactions · Community detection

1 Introduction

Taxes are divided into two types namely, direct taxes and indirect taxes. The major difference between these two is the way in which they are collected. Direct taxes are collected from individuals and corporations. Income tax and gift tax are examples of direct taxes. Indirect taxes are imposed on the goods and services consumed. In this work, we work towards detecting evasion prevailing in the indirect taxation system. Value-added Tax (VAT) [26], and Goods and Services Tax (GST) [5] are indirect taxes. They are collected by a third party (*eg.*, shop keeper) from the consumer who purchases the goods. Finally, it is the consumer who would have to bear the burden of the tax payment.

Recent tax reforms in developing countries opted indirect taxation method to expand their tax base. Determining the "point of levy" is an involved task in indirect taxes. A simple approach is to levy and collect the tax at a single point

K. Chen et al. (Eds.): BigData 2019, LNCS 11514, pp. 95–109, 2019.
https://doi.org/10.1007/978-3-030-23551-2_7

in the value chain, for example, the point of final consumption. The retail sales tax (RST) in the United States of America is an example. Single point of the levy is easy for administering but it has a few flaws. Many developing countries have a high concentration of informal economic activities at the consumption points. For example, the market share of informal economic activities in India is almost 50%. In these countries, there is a major risk of loosing out tax at the final consumption point. Sales tax can be sidestepped by taking the goods out of the value chain right at the onset. This will result in the creation of a parallel economy by keeping a major part of the value chain outside the regulatory authority's watch. One approach towards handling this problem is by following a multipoint taxation system, such as the Value-added Tax (VAT) and the Goods and Services Tax (GST) [5]. Goods and Services Tax, which is implemented in India from July 2017, is a comprehensive, multi-stage, destination-based tax that is levied on every value addition. This tax has replaced many indirect taxes that were previously existed in India.

1.1 Multipoint Taxation System (VAT and GST)

In this system, the tax is levied incrementally in each stage of the production depending upon the value added to goods in the corresponding production phase. Tax is levied at each phase of the production, such that tax paid on purchases(input tax) will be given as set-off for that tax levied on the sales (output tax) [8]. Figure 1 shows how the tax is collected incrementally in this system.

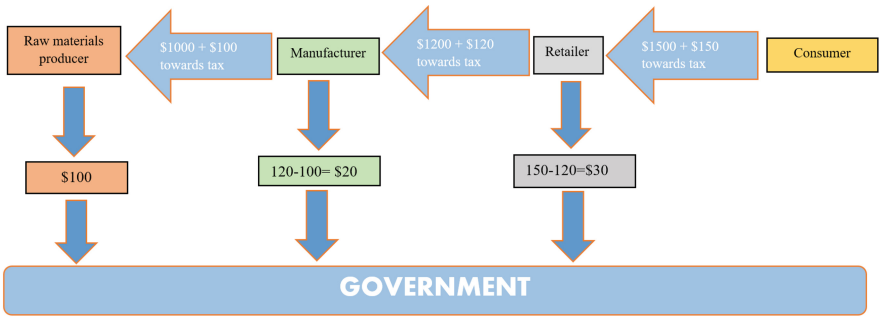

Fig. 1. Multipoint taxation system

- In this example, the manufacturer purchases some raw material of value 1000$ from the raw material dealer, by paying 100$ as tax at 10% tax-rate. The raw materials dealer remits to the government the tax amount that he has collected.
- Then the retailer purchases the processed goods from the manufacturer for, say, 1200$. An amount of 120$ is then paid to the manufacturer as a tax. The manufacturer pays the government the difference between the tax he had collected

(from the retailer) and the tax he has paid (to the raw materials producer) (120$ − 100$ = 20$).
– The consumer then buys the finished goods from the retailer for 1500$ by paying a tax of 150$. By following the same argument as given in the previous steps, the retailer pays 30$(i.e., 150$ − 120$) to the government.

It can be easily calculated that the total tax received by the government is 150$, and it is indirectly paid completely by the consumer of goods. Hence, raw material dealer, manufacturer, and retailers are representatives of the government to collect the tax.

This method ensures market-driven checks and balances or compliance regime which is difficult to achieve. At each node in the value chain, the purchaser and seller duo would have contradicting goals towards their tax liabilities. The seller tries to understate his sales while the purchaser tries to overstate purchases. This contradicting approach ensures market-driven checks and balances.

1.2 Tax Evasion Methods in Multipoint Taxation System

Circular Trading in GST and VAT: In GST and VAT, the market-driven checks and balances did not work in an expected manner. In the majority of tax evasion cases, business dealers, in their monthly tax-returns, deliberately manipulate their actual business transactions motivated by the amount of profit gained by evading tax. Invoice trading is a method to evade tax [2], where a dealer sells their goods to the end user without issuing the invoice but collecting the tax. Later, he/she issues a fake invoice to a third party, who uses it towards increasing their input tax credit. This will minimize the amount of tax they have to pay in the form of cash(the difference between the tax they collected at the time of sales and the tax they paid at the time of purchases) to the government. These tax manipulations can be spotted by the tax enforcement officials. To hide these manipulations, malicious dealers create a well-entrenched "racket," where a large number of bogus firms(shell firms) are created to manipulate the title of goods in the first place and then follow it up by making fake transactions among them to outwit easy systemic detection. Malicious dealers, show huge fake sales and purchases among malicious dealers and dummy dealers(shell firms) without any significant *value-added* as given in Fig. 2.

In Fig. 2, illegitimate transactions corresponding to fake invoices (invoice trading) are shown using red-lines. These are from dealers, x to q, x to z and q to z. Dealers q and z use these fake invoices towards minimizing their tax liability. With the motivation of confusing the tax enforcement officers, these dealers superimpose several fake transactions(dummy transactions) on these illegitimate transactions, which are shown using gray lines. Note that these dealers superimpose fake transactions such that the tax liability of any of the dealer due to the fake transactions is zero, *i.e.*, an amount of tax paid on the fake purchases is equal to the amount of tax collected on the fake sales.

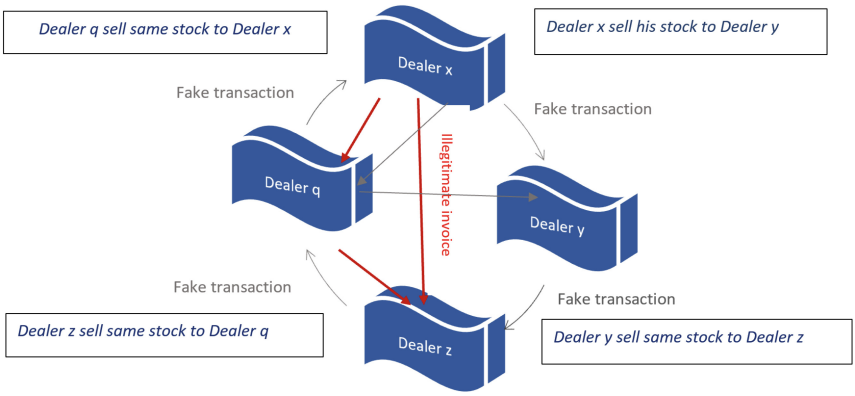

Fig. 2. Circular flow of sales/purchases

Since the value-addition due to the fake transactions is equal to zero, they do not pay any tax on these fake transactions, but, rather they create confusion to the tax officials about the illegitimate transactions. It is important to note that there is a huge amount of fake sales and purchases transactions among malicious and dummy dealers when compared to genuine sales and purchases transactions with the others. This type of technique used to evade tax is known as *circular trading* [9,10,21]. Hence the malicious dealers complicate the process of detecting their illegitimate transactions (invoice trading).

Carousel Fraud: Carousel fraud is a method of stealing public money by exploiting the VAT-free trade arrangements between European Union member countries. An organized crime groups will import goods from another country, then sell them by charging VAT to the customer but absconding with VAT instead of passing it to the government. To make this process undetectable, these groups buy and sell the goods multiple number of times between bogus companies before the final transaction where the VAT is stolen [4,18].

Carousel fraud and circular trading have a lot of characteristics in common. The solutions for circular trading can be extended to carousel fraud. In this paper, we work on circular trading.

1.3 Motivation for This Work

Manually, it is impractical for the tax officials to detect illegitimate transactions in circular trading due to the enormous size of the tax department's database, complicated sequences of sales and purchases transactions by the malicious dealers, the unknown identity of the traders doing these manipulations, *etc.* These challenges call for sophisticated big data and graph theoretic techniques. We used the *RHadoop* framework [6] for implementation.

The following gives a brief account of the paper structure. In Sect. 2, we describe several existing approaches that are used to perform cluster analysis on problems similar to that of ours. In Sect. 3, the problem is formulated as detecting communities in social networks and removing cycles created by fake transactions. In Sect. 4, we outline the experimental setup and results obtained from this work. We implemented these algorithms for the Commercial Taxes Department, Government of Telangana, India.

2 Related Work

Circular trading is a notorious problem in stock markets. In [21], Palshikar et al. proposed a highly customized algorithm for identifying colluding sets in stock trading. In [27], Wang, et al. proposed an algorithm to identify colluding sets in the instrument of future markets. In [13], Islam, et al. had given an algorithm for identifying collusion sets and cross trading collusion sets.

In [20], Nigrini et al. suggested statistical methods which can be used in the initial stages of the auditing. These techniques are based on Benford's Law, a unique characteristic of tabulated numbers. This law gives the expected probability of the digits in tabulated data. In [1], Arben Asllani et al. proposed a method that can be used by charted accountants to detect accounting fraud.

In [16], Klymko et al. have given an undirected edge weighting method based on directed triangles to detect communities in directed networks. They proposed a new measure on the quality of the communities in social networks depending on the number of 3-cycles that are span across communities. They showed that the resulting communities have fewer 3-cycles cuts. In [14,23], the author showed the significance of triangles in community detection in an undirected networks. In [15], Khadivi et al. showed that proper assignment of weights to the edges of a social network could improve community detection. They used this weighting as an initial step for the Newman greedy modularity optimization algorithm. In [17], the authors have proposed a method which can identify classification rules to detect fraudulent samples. They discovered spatial relationships of fraud and non-fraud financial statements. In [12], the authors have proposed a clustering based data mining algorithm to find outliers in taxation data. In [11], the authors have used clustering algorithms to identify a group of taxpayers, and then they have used several classification models to detect a potential user of false invoices in a given year.

In [6], Dean and Ghemawat explained MapReduce programming model for processing large data sets. In [3], Behera et al. had explained the implementation of random walk based graph clustering algorithm using Map-Reduce framework. In [25], Rajaraman et al. had given algorithms to handle massive data sets.

3 Problem Statement and Solution

It is impractical for the tax officials to detect illegitimate transactions in circular trading manually. Our objective is to design an algorithm to detect illegitimate

transactions and the set of a dealer doing these transactions. We follow the below four-step approach to solve the problem.

- **Step 1**: Construct an edge weighted directed graph from the way bill data base, where vertices correspond to dealers, and weights of directed edges are defined by the number of fake transactions, which are identified by Benford's analysis.
- **Step 2**: Convert this edge weighted directed graph into an edge weighted undirected graph.
- **Step 3**: Identify the groups of a dealer who perform excessive trade among themselves, as compared to the sales and purchases with other dealers. The problem is formulated as finding fraudulent communities in a social network.
- **Step 4**: Remove cycles formed by fake transactions within each group of these dealers.

3.1 Step 1: Construction of Sales Transaction Graphs

Waybill Database: Table 1 is a sample of a waybill data base. Each row corresponds to a sales transaction. Each row contains seller name, purchaser name, time of sales and value of sales.

Table 1. Waybill database

S.no	Seller	Purchaser	Time	Value
1	Tax Payer X	Tax Payer Y	2019/01/04/15:20	13000
2	Tax Payer Z	Tax Payer U	2019/01/04/17:00	19000
3	Tax Payer X	Tax Payer U	2019/01/05/19:00	15000
4	Tax Payer Y	Tax Payer Z	2019/01/05/17:00	15000
5	Tax Payer Z	Tax Payer X	2019/01/05/15:30	13000

The actual database contains many more details like type of goods, the rate of tax, the quantity of goods, vehicle used for transporting the goods, vehicle number, transporter name, invoice number, UOM (unit-of-measure), inserted date, etc. The data we had taken contains several million rows.

Benford Analysis: Benford's law, which is also known as the first digit law, is a statistical technique for fraud detection [1,7,20]. This law intrigued mathematicians for over a century. This law gives the probability of the leading digit in a naturally occurring numeral data.

The Benford's law states that for any numerical data with a distribution of numbers spanning several orders of magnitude(an order of magnitude is an approximate measure of the number of digits that a number has in the

commonly-used base-ten number system), the probability of a number starting with the digit d is given by $log_{10}(1 + 1/d)$, where $d \in \{1, 2, ..., 9\}$.

Mean absolute deviation (MAD) is a statistical method which can be used to find whether the data's first digits follow the probability distribution given by Benford's law. Mean absolute deviation $MAD = \sum_{j=1}^{m}(OP_j - EP_j)/m$, where OP_j is the observed probability of j^{th} bin, EP_j is the expected probability of j^{th} bin, and m is the total number of bins (in this case it is equal to 9). Based on the MAD value, we can find the conformity between expected probability and observed probability as given below [19].

- MAD value between 0.000 to 0.004 says "Close conformity"
- MAD between 0.004 to 0.008 says "Acceptable conformity"
- MAD between 0.008 to 0.012 says "Marginally acceptable conformity"
- MAD greater than 0.012 says "Nonconformity"

Sales Transaction Graph: We use waybill database to construct an edge weighted directed social network denoted by $G_d = (V_d, E_d)$, where V_d is the vertex set (each dealer corresponds to a vertex), and E_d denotes the set of weighted directed edges. We name this social network as *sales transaction graph*. Below we propose a method to assign weights to the edges.

Let m be the number of sales transactions in the waybill database from dealer vertex x to dealer vertex y and $v_1, v_2, v_3, \ldots, v_m$ be values of these sales. Let $\beta(xy)$ be the MAD value of the first digit Benford's analysis on $v_1, v_2, v_3, \ldots, v_m$. Based on the value of $\beta(xy)$, we can establish the conformity between expected and observed distribution.

The weight $w(xy)$ of the edge from vertex x to vertex y in graph G_d is given by $w(xy) = (m * \sum_{i=1}^{m} v_i)/(m + \sum_{i=1}^{m} v_i) * e^{1000*\beta(xy)}$. Note that lesser edge weights are assigned for the edges with less number of transactions or less sum of the values of the transactions [24]. The weight of the edge xy increases exponentially with $\beta(xy)$, i.e., more weight is assigned for a lesser conformity between expected distribution and observed distribution.

3.2 Step 2: Construction of Weighted Undirected Graph

Majority of work in community detection has been done on undirected graphs. In this paper, we propose a method to convert an edge weighted directed graph into an edge weighted undirected graph.

There are several metrics to measure the quality of a community. One major idea is that flow tends to stay within the community. Hence, cycles in a graph play an important role in community detection. In detecting communities in circular trading, 2-cycles and 3-cycles play an important role. We propose a weighting scheme to turn an edge weighted directed graph to an edge weighted undirected graph. The weight given for an edge is based on triangles and two cycles in which this edge is involved [15,16,23].

In the following, we will explain how to construct an edge weighted undirected graph $G_u = (V_u, E_u)$ from an edge weighted directed graph G_d described in Subsect. 3.1. Let $C = (a, b, c)$ be a cycle in G_d. Cycle C can be any one of the four types of cycles shown in Fig. 3. The weight of cycle C is defined as follows.

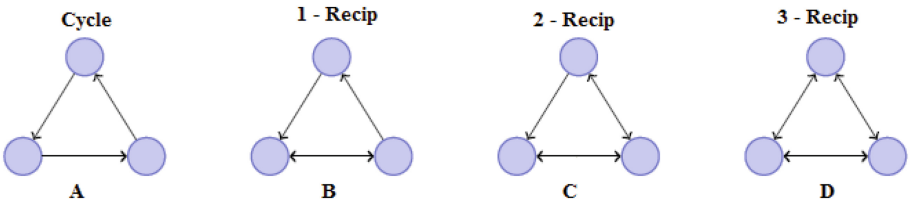

Fig. 3. Different types of 3-Cycles

- If it is of *Type a*, then the weight of the cycle C is given by $W(C) = min\{w(ab), w(bc), w(ca)\} * 1$
- If it is of *Type b*, then the weight of the cycle C is given by $W(C) = min\{w(ab), w(bc), w(ca)\} * 1.5$
- If it is of *Type c*, then the weight of the cycle C is given by $W(C) = min\{w(ab), w(bc), w(ca)\} * 1.75$
- If it is of *Type d*, then the weight of the cycle C is given by $W(C) = min\{w(ab), w(bc), w(ca)\} * 1.875$

The weights 1, 1.5, 1.75 and 1.875 are given to 3-cycles of types a, b, c, d respectively. These are chosen empirically based on the clustering performance. The number of reciprocal edges in a triangle conveys the strength of circular trading. Hence, we gave more weight to a 3-cycle with more number of reciprocal edges.

Suppose directed edge ab is in cycles $C_1, C_2, \ldots, C_m$ and directed edge ba is in cycles $D_1, D_2, \ldots, D_n$. Then the weight of an undirected edge ab in graph G_u is given by the maximum among the following three values

- $max\{W(C_1), W(C_2), \ldots, W(C_m)\}$
- $max\{W(D_1), W(D_2), \ldots, W(D_n)\}$
- $min\{w(ab), w(ba)\} * 2$

3.3 Step 3: Community Detection

We use WalkTrap algorithm to detect communities. WalkTrap algorithm is a hierarchical agglomerate clustering algorithm and it uses a distance measure based on random walks [22]. It is based on the assumption that a random walker would spend a longer time inside a strong community due to the high density

of edges within the community. This algorithm measures the similarity between vertices and between communities by defining a distance between them. This distance measure is calculated from the probabilities that the random walker moves from one vertex to another in a fixed number of steps.

Distance Between Communities. Let us consider random walks of a given length t on graph G. Let p_{ij}^t be the probability of reaching vertex j from vertex i in a random walk of length t. Value of t should be large enough to capture the community structure of G but not too large to reach a stationary distribution. Generally, the value of t is between three and six. The basic idea behind this algorithm is two vertices of the same community tend to *see* all the other vertices in the same way. Thus if vertices i and j are in the same community, we can expect that $\forall k,\ p^t ik \cong p^t jk$. Then the distance between vertices i and j can be defined as $\sqrt{(\sum_{k=1}^{n} \frac{(p_{ik}^t - p_{jk}^t)^2}{d(k)})}$, where $d(k)$ is the degree of vertex k [22]. One can generalize the distance between vertices to a distance between communities in a straightforward way.

3.4 Step 4: Removing Cycles in Each Cluster

Consider any community (cluster) C given by the community detection algorithm. Using the waybill database explained earlier, we construct a directed edge-labeled multi-graph called *sales and purchase graph*, denoted by $G_{sp} = (U, E, \gamma)$, where U is the set of vertices (each vertex corresponds to a dealer in C), E is the set of labeled directed edges (an edge from vertex x to vertex y corresponds to a sales transaction from x to y) and γ is the function that associates a 2-tuple for each labeled edge, where the first element of the tuple is the time of sales of this transaction and the second element is the value of sales of this transaction.

Following are few notations we use. Note that each edge in the graph has two parameters, one is the time of sales and the other is the value of sales. The *end_time* of a cycle is defined as the time of most recent sales transaction among all the sales transactions corresponding to the edges in the given cycle. The *start_time* of a cycle is defined as the time of least recent transaction among all the sales transactions corresponding to the edges in the given cycle. The *time_gap* of a cycle is defined as the difference between *end_time* and *start_time*. The *maxval* of a cycle is defined as the maximum value among values of all the transactions corresponding to the edges in the given cycle. The *minval* of a cycle is defined as the minimum value among values of the transactions corresponding to the edges in the given cycle. The *valgap* of a cycle is defined as the difference between *maxval* and *minval*. Let the *trust score* of a cycle is defined as *time_gap* * *valgap*.

From in-depth research by taxation authorities, it is observed that *time_gap* and *valgap* of any fake sales cycle are very small, which means the *trust score* of any fake cycle is small. Our motive is to remove all fake cycles from the *sales and purchase graph*. Then the remaining graph will be a directed acyclic graph (DAG). Note that the resultant directed acyclic graph contains all suspicious

transactions. This makes fraud detection process simpler which allows us to do a deeper analysis on suspicious transactions to identify tax evaders. Below we give a brief sketch of the fake cycles removal algorithm.

1. Select a cycle D in sales and purchase graph G_{sp} with the following conditions:
   - *Condition 1: end_time* of D is minimum among all the cycles in G_{sp}
   - *Condition 2:* With respect to the condition one, *trust score* of D is minimum
   - *Condition 3:* With respect to the condition two, *length* of D is minimum
2. Let y be the minimum of the values of sales of all the edges in D. Subtract y from values of sales of all edges in D.
3. Remove any edge from D whose value of sales becomes zero.
4. Repeat steps one to three, as long as G_{sp} contains a cycle.

3.5 Algorithms

Detecting and Managing Outliers: According to Benford's analysis, the probability of nine occurring as the first digit is 0.046 [19]. We need at least twenty-two transactions between any pair of dealers to get a valid Benford's score. As part of data cleansing, we remove sales transactions between pairs of dealers (vertices) if the number of sales transactions between them is less than twenty-two. If the value of any sales transaction in waybill database is more than *third quantile plus 1.5 times the inter-quantile range of values of sales transactions*, then replace the value of this sales transactions by *third quantile plus 1.5 times the inter-quantile range of values of sales transactions* [2].

Algorithms: Algorithm 1 is a community detection algorithm. First, we apply this algorithm to detect communities. Removing outliers and construction of directed graph are highly time consuming operations in this algorithm due to millions of purchase and sales transactions. These operations are parallelized. We used Map-Reduce framework to implement these operations. Later we apply Algorithm 2 to identify illegitimate transactions. Note that both algorithms are polynomial time algorithms.

4 Case Study

4.1 Experimental Setup

We used the *R* programming language for data mining and Hadoop framework for storing data. We used the *RHadoop* open source analytics solution to integrate *R* programming language with *Hadoop*.

Data: WayBill Data
Result: Set of Communities

Perform outlier cleansing;
`# This is explained in 3.5.`;

Construct a directed graph G_d;
`# This is explained in 3.1. Note that Benford's analysis has to be performed on the values of sales transaction before outlier cleansing.`;

Construct an undirected graph G_u;
`# This is explained in 3.2`;

Find communities in G_u using WalkTrap algorithm;
`# If any community is bigger than eight vertices, perform sub-community detection on this community`;
}

Algorithm 1. Community detection algorithm

4.2 Identifying Communities

In our data set, there are 0.6 million dealers. Size of our data set is 1.5 TB. Figure 4, shows the business among some of these dealers. We applied the Algorithm 1 on this data set and obtained several communities, which are doing heavy circular trade. Figure 5, shows a few communities obtained. We used two measures namely modularity and coverage to validate the clustering. Modularity and coverage are 0.74 and 0.82 respectively.

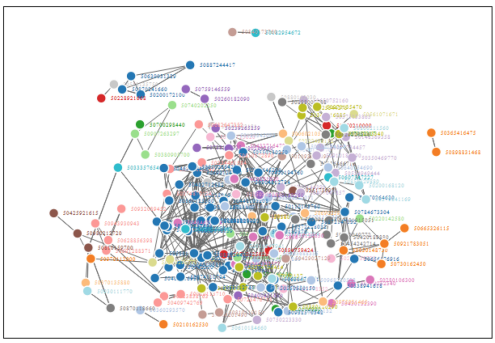

Fig. 4. Complex network of sales and purchases

Data: sales and purchase graph G_s
Result: Forest G_t, which is obtained by removing all cycles in G_s

G_t = Edgeless graph whose vertex set is $V(G_s)$;

Let $l_1, l_2, \ldots, l_m$ be a sequence of all edges in G_s ordered by non decreasing order of time of sales;

for $i = 1 \ldots m$ **do**
  insert the edge l_i in the graph G_t;
  while *(G_t contains cycle)* **do**
    Assume that the edge l_i is from vertex b to vertex a in G_t;
    Let $P_1, P_2, \ldots, P_k$ be set of the path from a to b in G_t;
    Let sp_i, vg_i be *time_gap* and *valgap* of cycle C_i formed by path P_i along with the edge ba, for $1 \leq i \leq k$;
    Let $spdiff = \max\{sp_1, sp_2, \ldots, sp_k\}\text{-}\min\{sp_1, sp_2, \ldots, sp_k\}$;
    Let $valgapdiff = \max\{vg_1, vg_2, \ldots, vg_k\}\text{-}\min\{vg_1, vg_2, \ldots, vg_k\}$;
    Let *normalised trust score* of cycle C_i be $(sp_i/spdiff) * (vg_i/valgapdiff)$ for $1 \leq i \leq k$;
    Let C_j be a cycle, where $1 \leq j \leq k$, such that *normalized trust score* is minimum;
     `# This cycle can be identified in polynomial time`;
    Let p be the minimum among the price of sales of all edges in C_j;
    Subtract p from the price of sales of all edges in C_j;
    Remove all edge from G_t whose price of sales is zero;
  end
end

Algorithm 2. Cycle removal algorithm

4.3 Identifying Illegitimate Transactions

We had taken one community with four dealers which is shown in Fig. 6. These four dealers are doing heavy circular trade among themselves. Their sales, purchase and tax details are shown in Fig. 7. Total tax paid by these four dealers is Indian rupees 0.03 million which are shown in column seven. The tax they collected on sales(output tax) is Indian rupees 367 million as shown in column six. They set-off this entire tax collected with the tax they paid on purchases(input tax) which is shown in column four. In genuine Iron and Steel, the business ratio between the input tax and output tax will be less than 0.95, but here it is almost one. We applied Algorithm 2 on this community to remove fake cycles and identify illegitimate transactions. When the tax authorities physically visited the premises of these companies, they identified that these are shell companies.

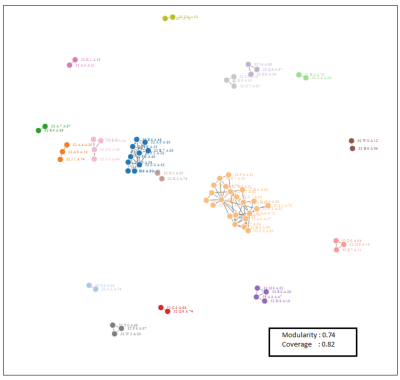

Fig. 5. Experimental result

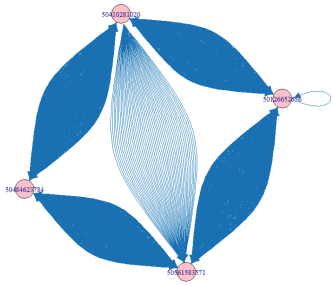

Fig. 6. Cluster of four dealers

S No	Dealer	Purcahse Amount	Input Tax Credit	Sales Amount	Output Tax	Tax Payment
1	2	3	4	5	6	7
1	50126652656	1760.68	88.03	1731.39	86.57	0.00
2	50410281020	1998.65	100.23	2021.42	101.21	0.00
3	50484623784	1711.67	85.58	1712.72	85.64	0.03
4	50561583571	1996.01	99.80	1902.17	95.11	0.00
TOTAL		7467.01	373.64	7367.7	368.53	0.03

Fig. 7. Business details

5 Conclusion

Here we studied a widely practiced tax evasion method in GST called *circular trading*. *Circular trading* is a tax evasion practice where a set of malicious dealers do heavy fake sales and purchase transactions among themselves that go around in a circular manner in a very short time-duration without any meaningful *value-addition*. They practice this technique to hide illegitimate transactions. We addressed the problem of identifying the cluster of dealers who do excessive fake trade among themselves and illegitimate transactions performed by them. We implemented this technique using *RHadoop* big data framework for the Commercial Taxes Department, Government of Telangana, India. Our results are helping the tax authorities to effortlessly identify illegitimate transactions and take legal action against those who are doing these transactions. As future work, we plan to work on developing sophisticated algorithms that detect colluding communities by exploiting the different patterns made by the fraudulent dealers.

Acknowledgment. We would like to express our deep thanks towards the government of Telangana, India, for allowing us to use the Commercial Taxes Data set and giving us constant encouragement and financial support. This work has been supported by Visvesvaraya Ph.D. Scheme for Electronics and IT, Media Lab Asia, grant number EE/2015-16/023/MLB/MZAK/0176.

References

1. Arben Asllani, M.N.: Using Benford's law for fraud detection in accounting practices. J. Soc. Sci. Stud. **1**, 129–143 (2014)
2. Baesens, B., Vlasselaer, V., Verbeke, W. (eds.): Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection. Wiley, Hoboken (2015). ISBN 978-1-119-13312-4
3. Behera, R., Rath, S., Misra, S., Damasevicius, R., Maskeliunas, R.: Large scale community detection using a small world model. Appl. Sci. **7**, 1173 (2017)
4. Borselli, F., Fedeli, S., Giurato, L.: Digital VAT carousel fraud: a new boundary for criminality. Tax Notes International (2015)
5. Dani, S.: A research paper on an impact of goods and service tax (GST) on indian economy. Bus. Econ. J. **7**, 264 (2016). ISSN 2151–6219
6. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. In: Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation, OSDI 2004, vol. 6, p. 10. USENIX Association, Berkeley (2004). http://dl.acm.org/citation.cfm?id=1251254.1251264
7. Durtschi, C., Hillison, W., Pacini, C.: The effective use of Benford's law to assist in detecting fraud in accounting data. J. Forensic Account. **V**, 17–34 (2004)
8. Dutta, R., Kumar, B.: Value added tax scams and introduction of the goods and services tax. Econ. Polit. Wkly. **53**(44) (2018)
9. Franke, M., Hoser, B., Schröder, J.: On the analysis of irregular stock market trading behavior. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.) Data Analysis, Machine Learning and Applications, pp. 355–362. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-78246-9_42. ISBN 978-3-540-78239-1
10. Golmohammadi, K., Zaiane, O., Díaz, D.: Detecting stock market manipulation using supervised learning algorithms. In: Data Science and Advanced Analytics, pp. 435–441. IEEE, November 2014. http://ieeexplore.ieee.org/document/7058109/, ISBN 978-1-4799-6991-3
11. González, P.C., Velásquez, J.D.: Characterization and detection of taxpayers with false invoices using data mining techniques. Expert Syst. Appl. **40**(5), 1427–1436 (2013)
12. Huang, S.Y., Tsaih, R.H., Yu, F.: Topological pattern discovery and feature extraction for fraudulent financial reporting. Expert Syst. Appl. **41**(9), 4360–4372 (2014)
13. Islam, N., Rafizul Haque, S., Masudul Alam, K., Tarikuzzaman, M.: An approach to improve collusion set detection using MCL algorithm. In: Computers and Information Technology, pp. 237–242. IEEE, December 2009. http://ieeexplore.ieee.org/abstract/document/5407133/, ISBN 978-1-4244-6284-1
14. Berry, J.W., Hendrickson, B., LaViolette, R.A., Phillips, C.A.: Tolerating the community detection resolution limit with edge weighting. Phys. Rev. E **83**, 056119 (2011)

15. Khadivi, A., Ajdari Rad, A., Hasler, M.: Network community-detection enhancement by proper weighting. Phys. Rev. E **83**, 046104 (2011). https://doi.org/10.1103/PhysRevE.83.046104
16. Klymko, C., Gleich, D.F., Kolda, T.G.: Using triangles to improve community detection in directed networks. ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference, Stanford University abs/1404.5874 (2014)
17. Liu, B., Xu, G., Xu, Q., Zhang, N.: Outlier detection data mining of tax based on cluster. Phys. Procedia **33**(44), 1689–1694 (2012)
18. Frunza, M.-C.: Aftermath of the VAT fraud on carbon emissions markets. J. Financ. Crime **20** (2013)
19. Mark Nigrini, J.T.W. (ed.): Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection. Wiley, Hoboken (2012). ISBN 978-1-118-15285-0
20. Nigrini, M.J., Mittermaier, L.J.: The use of Benford's law as an aid in analytical procedures. Audit.: J. Pract. Theory **41**, 52 (1997)
21. Palshikar, G., Apte, M.: Collusion set detection using graph clustering. Data Min. Knowl. Discov. **16**, 135–164 (2008). https://doi.org/10.1007/s10618-007-0076-8. ISSN 1384–5810
22. Pons, P., Latapy, M.: Computing communities in large networks using random walks. In: Yolum, I., Güngör, T., Gürgen, F., Özturan, C. (eds.) ISCIS 2005. LNCS, vol. 3733, pp. 284–293. Springer, Heidelberg (2005). https://doi.org/10.1007/11569596_31
23. Prat-Pérez, A., Dominguez-Sal, D., Brunat, J.M., Larriba-Pey, J.L.: Shaping communities out of triangles. In: ACM International Conference Proceeding Series, July 2012
24. Jarvis, R.A., Patrick, E.A.: Clustering using a similarity measure based on shared nearest neighbors. IEEE Trans. Comput. **C–22**(11), 1025–1034 (1973)
25. Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press, New York (2011)
26. Schenk, A., Oldman, O. (eds.): Value Added Tax: A Comparative Approach. Cambridge University Press, Cambridge (2007). ISBN 978-1107617629
27. Wang, J., Zhou, S., Guan, J.: Detecting potential collusive cliques in futures markets based on trading behaviors from real data. Neurocomputing **92**, 44–53 (2012)