Link prediction techniques to handle tax evasion

Jithin Mathews
Indian Institute of Technology Hyderabad
Hyderabad, Telangana, India
cs15resch11004@iith.ac.in

Priya Mehta
Indian Institute of Technology Hyderabad
Hyderabad, Telangana, India
cs15resch11007@iith.ac.in

Suryamukhi K
Indian Institute of Technology Hyderabad
Hyderabad, Telangana, India
cs17m19p100001@iith.ac.in

Sobhan Babu
Indian Institute of Technology Hyderabad
Hyderabad, Telangana, India
sobhan@iith.ac.in

ABSTRACT

Circular trading of goods is a carefully designed scam ubiquitous among fraudulent business dealers all around the world. Dealers involved in this scheme create an artificial trading network by issuing doctored sales-invoices amongst themselves without any movement of goods. In practice, it is observed that almost all cases of circular trade involve two or three dealers. Here, we work towards predicting circular trade involving three dealers. For the same, we built four different classification models consisting of feature variables tailored for predicting any plausible circular trade amongst three dealers. In particular, the logistic regression model gave the best performance among all the four different models with a prediction accuracy of 80%. Interestingly, we observe that a feature variable formed by using the personalised PageRank technique significantly improves the model over the state of the art link prediction variables. Predicting a future circular trade from a huge network of sales-transactions data is of significant importance to the tax enforcement officers. In addition to automating the process of detecting circular trading, which is manually impossible, this model helps them to target on a set of plausible evaders and take appropriate preventive measures. This model have been developed for the Commercial Taxes Department, Government of Telangana, India, using their first two quarter's tax returns dataset.

CCS CONCEPTS

• **Applied computing** → **Economics**; • **Computing methodologies** → *Supervised learning by classification.*

KEYWORDS

link prediction, goods and services tax, circular trading, tax evasion, forensic accounting, logistic regression, PageRank algorithm

ACM Reference Format:
Jithin Mathews, Priya Mehta, Suryamukhi K, and Sobhan Babu. 2021. Link prediction techniques to handle tax evasion. In *8th ACM IKDD CODS and 26th COMAD (CODS COMAD 2021), January 2–4, 2021, Bangalore, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3430984.3430998

CODS COMAD 2021, January 2–4, 2021, Bangalore, India

ACM ISBN 978-1-4503-8817-7/21/01…$15.00
https://doi.org/10.1145/3430984.3430998

1 INTRODUCTION

Circular trading is a fraudulent scheme where a business dealer creates an artificial trading loop by issuing fake sales invoices (bill trading [6]) to a small group of dealers without any actual supply of goods. In order to avoid any tax liability due to these fake transactions, the dealers involved in this scam make sure that there is no value addition due to such transactions. It is observed that at least one of the company in the ring of traders involved in circular trade would be bogus, and is termed as *shell*[8] companies in the literature. A circular trade transaction is generally doctored in such a way that its legal ramifications will be upon the *shell* companies while the other dealers, who actually planned the fraud, can excuse themselves as innocent traders who got caught upon the scam[8]. Therefore, these *shell* companies serves both as a conduit and finally as exit points for the actual culprits who are involved in the circular trade when an investigation boils up.

1.1 Purpose of Circular Trading

- To increase the turnover of a business.
- To avail bigger loans from banks & NBFCs (Non-Banking Financial Corporations).
- To bring black money into the system.
- To increase valuation of companies.

In a nutshell, circular trading refers to the transaction of selling and buying of goods in a loop (without actual movement of goods) through *shell* companies to inflate business turnover[7].

1.2 An Illustration for Circular Trading

Figure 1 illustrates a simple scenario of *circular trading* involving 3 business dealers, *viz*., A, B and C. Initially, dealer A sells some goods to dealer B, dealer B then sells the same kind of goods to dealer C, and finally when dealer A buys the same type of goods from dealer C, the cycle is complete. As it can be observed from Figure 1, the $Value$ of goods transferred is almost the same in all the three transactions. These type of flow of goods in the trade network, in which, the same kind of goods is cycled around in a circular manner is not expected for certain kinds of goods and commodities in the market. Therefore, the presence of cycles become an indicator for fraudulent transactions in such instances.

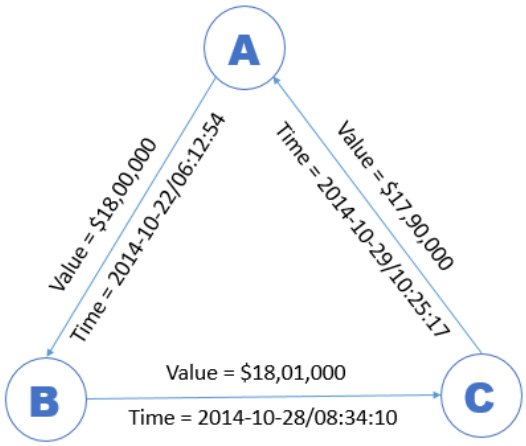

Figure 1: Circular trading

1.3 Observations from Circular Trading

$$Tax\ liable = (Output\ tax - Input\ tax) \quad (1)$$

In equation 1, *Tax liable* is the amount of tax a dealer is liable to pay to the government. *Output tax* refers to sales tax, *i.e.*, the tax collected by the dealer (from the buyer of goods) by selling goods. *Input tax* refers to the tax given by the dealer (given to the seller of goods) while buying goods.

Since there is no value addition on the fake transactions used for circular trading, the net *Tax liable* amount due to these transactions is near to zero. As the market price of goods can vary from one day to the other, to make the net *tax liable* amount to zero, dealers are limited to one option:

- Raise the invoices in such a way that all the transactions happen in a very short period of time.

If they do not raise the invoices such that the transactions happen in a very short period of time, then they are left with no other choice to mask their crime. Suppose they raise the invoices in a gap of more than a day. Then the price of the goods can vary due to market fluctuations from one day to the other. So dealers are left with the only option to increase the *Input tax*, or, to decrease the *Output tax*, to make the net *Tax liable* amount to zero. This can be done only by increasing the amount of goods purchased, or, by decreasing the amount of goods sold, respectively. In both the cases, after buying and selling of goods, ideally there should be some quantity of the same goods remaining in the warehouse of the dealer. However, in case of a warehouse audit, they will have no account for this remaining goods since there are no actual goods being traded in these doctored transactions.

Hence it is easy for the taxation authorities to detect these transactions when dealers make adjustment in the quantity of goods purchased or sold. Therefore, in the invoices, the dealers performing circular trading are forced to show that the transactions happen in a very short period of time, preferably within few hours. Table 1 gives the percentages of cycles formed in the dataset within the duration of one day, two days, three days, four days and fourth day to the end of the month. Notice that the link forming the cycle, *i.e.*

Table 1: Percentages of cycles

	Cycle formation period in days				
	1	2	3	4	4-31
%	97.26	1.15	0.69	0.48	0.42

Table 2: Percentages of cycles of different length

	Cycle length			
	2	3	4	>4
%	57.91	41.78	0.19	0.09

the last edge formed in the cycle, can be assigned as the last edge only to that particular cycle. The percentages are relevant only up to the end of the month since the dealers file their tax return statements on a monthly basis.

1.4 Our Contribution

Palshikar et al.[16] proposed certain approaches for detecting circular trading in stock markets. They devised a graph clustering algorithm customized for identifying collusion parties in the stock market trading. In this work, we are trying to predict circular trading before it happens. During data exploratory analysis, it is observed by the domain experts that almost all the cases of circular trade involves two or three dealers. We note that, on an average, we observed 3 cycles of length greater than three for every thousand cycles of length two or three. Interestingly, as observed in [16], this is not the case for circular trading in stock markets where a significantly large number of traders collude with each other. In the taxation realm, the number of dealers involved in circular trading are limited due to the complexities involved in maintaining a large group of accomplices, and the fraudsters naturally incline to keep the profit among a trusted group of loyal acquaintances which usually would be very small in number. Table 2 gives a glimpse on the frequency of cycles of different lengths present in the data set.

Here, we work towards predicting circular trade involving three dealers. We built four binary classifier models for the same. It consists of feature variables customised by entwining the information gathered from domain experts (tax enforcement officers) and the techniques used in link prediction from the literature. In particular, the logistic regression model perform the best. This model predicts whether a link, that causes the creation of a three cycle, would be formed in future between two dealers, with an accuracy of 80%. The classification model developed here is yet to be used by the taxation officials. Our model uses labelled data containing the fraudulent links that causes three cycles. Note that the labelled data pertains to the dealers trading certain set of sensitive goods (shared to us by the tax-officials) for which the buying and selling of same goods in a loop is not an expected pattern.

Predicting a circular trade before it happens is a game changer for the taxation department. Currently they are employing a post-mortem approach in dealing with the fraud. It is both very time consuming and less worthy in comparison with the efforts and logistics spend on proving the scam in court. Alternatively, it opens up a plethora of resources in the tool-kit of tax officials, if they are given in advance a plausible list of traders who may perform

circular trade before the cycle is even formed. Giving warning notices can deter dealers from performing the scam. In case if some dealers still perform the transaction under the radar, that's the link causing the three cycle, investigations can be opened right from the outset. In a nutshell, this work helps the taxation officials in a significant manner to mitigate tax evasion due to circular trading, which otherwise was a cumbersome task of finding the needle in a haystack. Although there exists work on circular trading, to the best of our knowledge, this is the first work that addresses techniques to predict circular trade in the realm of taxation. Here, we have used original tax-returns data, containing invoice level details of every business transaction across the state of Telangana, India, for the first two quarters of the fiscal year 2015.

It is not difficult to deploy this model in the taxation system. If deployed the dataset will be streamed from the server to the host computer. The graph used in this situation will be a dynamic graph and not a static one, since new dealers can pop-up in the market. In the streaming paradigm, all the parameters used in the model will be updated with the addition of every new transaction. In this work, we have not included the streaming version, and keeps our motto towards describing the model used to predict the three cycles.

The rest of the paper is organised as follows. In Section 2, we discuss the previous relevant works. In Section 3, the regression model built to predict a plausible circular trade is described in detail along with all its feature variables. Experimental results, along with the model validation techniques are discussed in Section 4. In Section 5, the concluding remarks along with the future work plans are briefed.

2 RELATED WORK

Several significant works can be found in the literature that detects and combats fraudulent activities in a myriad of unrelated fields. Here we will go through a few notable ones.

Spectral clustering is employed in [19] for detecting under reported tax declarations in a city using silhouette value as the validation measure for cluster quality. In [15], authors introduced an approach which gives an estimate on the amount of tax lost by the government due to certain illegal activities performed by a particular set of suspicious dealers. In [3], authors showed that if clients are engaged with well-connected individual auditors then they have a comparatively lower effective tax rate. A technique using statistical methods are developed in [2] for detecting VAT evasion done by Kazakhstani business firms. In [17], a parallel tax fraud detection algorithm is introduced using Bayesian networks as the means for parallelization.

Several approaches are proposed for detecting circular trading in stock trading. In [16], Palshikar et al. used Dempster–Schafer theory to merge the colluding traders. In [24], neural networks are used for fraud detection, however it uses supervised training and thus works only for labelled data set. Wang et al. in [23] proposes an algorithm to identify colluding sets in the instrument of future markets. In [11], authors have used K-means clustering for finding malicious activities in a telecommunication company, and [5] uses data mining techniques for anomaly detection. In [9], authors used supervised algorithms in order to classify fraudster operations coupled with a clustering methodology.

In a separate work [10] the authors used a variation of the famous PageRank algorithm [4] called the Trust Rank for identifying and separating spam pages in web. In [20], credit card fraud detection problem is presented and the authors developed classification models based on Artificial Neural Network (ANN) and Logistic Regression(LR).

3 METHODOLOGY

3.1 Dataset

Table 3: Sales Database

No.	Seller's ID	Buyer's ID	Time	Value in $
1	a	b	2012/11/10/11:23:00	20000
2	c	d	2012/11/10/11:06:00	30000
3	d	b	2012/11/10/10:08:00	19000
4	m	n	2012/11/10/09:09:10	17000

The data used in this work is confidential and is provided by the commercial tax department of the state of Telangana, India. The raw data is more granular and huge in size, upto $4TB$, containing invoice level details of every business transaction across the state for the first six months in the financial year of 2015. The information relevant to this work is taken from the raw data after cleaning, and Table 3 shows a snapshot of this filtered dataset.

In this work, only the four parameters given in Table 3 are relevant and each row of the table represents a unique invoice. Parameter 'ID' represents a business dealer uniquely. Here, 'Seller's ID' and 'Buyer's ID' shows the direction of the flow of goods. Note that the flow of money is in the opposite direction, $i.e.$, from buyer to seller. 'Time' represents the time at which the transaction took place and the parameter 'Value' represents the amount of tax given by the buyer of the goods to the seller. For example, in the first row of Table 3 a dealer with ID a is selling goods to a dealer with ID b on 2012/11/10 at local time 11:23:00 AM. Dealer b gives a tax-amount of $\$20,000$ to dealer a.

We represent the transactions among the dealers using a weighted directed multi-graph $G = (V, \overrightarrow{E})$, where $V = (v_1, v_2, \ldots, v_n)$ is the vertex set containing the unique ID's of all the dealers in the transactions. The flow of money which is from the buyer to the seller is represented by the set of all directed edges $\overrightarrow{E}$. Note that a 2-tuple corresponding to the 'Value' and 'Time' attribute values, $(Value, Time)$, denote the weight on an edge in $\overrightarrow{E}$. Interestingly, all edges, including multiple edges, are uniquely identifiable using the combination of the three parameters, 'Seller's ID', 'Buyer's ID' and 'Time'. The reason for the same being the state's sales-purchases laws that restrict the issue of multiple invoices between two business dealers at exactly the same time. Note that the directed edges, $i.e.$ the relevant business transactions, analysed here are in the order of few millions.

Here we try to predict the formation of three cycles, which we call as "$triads$" [12] throughout the paper. In Figure 2 one can observe that both the edges $\overrightarrow{vo}$ and $\overrightarrow{ou}$ are already created. In this scenario, the formation of the directed edge $\overrightarrow{uv}$ (denoted by doted

lines) can create the triad $\overrightarrow{uvo}$. Hence we call the edge $\overrightarrow{uv}$ as a "*potential edge*" that may create the triad in the future.

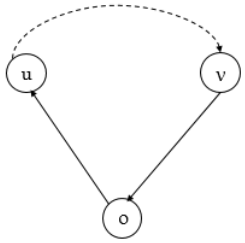

Figure 2: Triad $uvo^{->}$

3.2 Feature Variables

After building the classifier models, the five variables given in Table 4, *viz*., *JC*, *FR*, *TC*, *IV* and *PPR* are found to be statistically significant. These variables are chosen as feature variables for performing classification. The first four variables are derived from the local behavior of the vertices in the triad, while the fifth and the final feature variable focuses on the behavior of the entire connected component consisting of the triad. The last parameter given in Table 4, *viz*., 'Link' corresponds to the dependant variable that tells whether the potential edge $\overrightarrow{uv}$ will be formed in the future (Link=1) or not (Link=0). The five feature variables are described in detail in the subsection 3.2.1. Table 4 gives a snapshot of the data used for the same. Note that this data is derived from the Sales Database mentioned in Table 3.

Table 4: Potential-edge Database

No.	u	v	o	JC	FR	TC	IV	PPR	Link
1	a	b	c	0.1176	6.0000	4	0.7056	0.16789774	1
2	x	y	z	0.0714	2.0000	0	0.1428	0.01175115	0

In Table 4, every row corresponds to a potential edge. An edge $\overrightarrow{uv}$ is a potential edge if both the edges $\overrightarrow{vo}$ and $\overrightarrow{ou}$ are formed on the same day. This, as detailed in the introductory section, is due to the fact that the cycles in a circular trade are created by fraudsters in a short duration of time to withstand market fluctuations. Note that this data is taken from the Sales Database given in Table 3 without considering the first two months. The first two months are reserved for the data to get matured. This helps to form a data rich with different business patterns for different dealers, which is then tapped by combining the feature variables.

As it goes for most of the domain specific network datasets, straight forward application of the classical link prediction variables may not produce a good regression model. A Plentiful of research work has focused towards developing link prediction techniques, [14] gives a survey on them. This work is primarily motivated by [12] and [13] where the authors try to predict signed relationships, friend(+) or foe(−), between two parties based on the nature of their relationships with others in the social-network. Heuristic methods using Adamic/Adar index [1] and others [14] have been used from the beginning for link prediction problems. In [25], authors study a heuristic learning paradigm using Graph Neural Networks which learns the important heuristics for link prediction using local subgraphs.

3.2.1 *Construction of the Feature Variables.*

- Jaccard's Coefficient: If two dealers u and v collude each other to form the potential edge $\overrightarrow{uv}$ and hence forming the triad, then, it is normal to think that they have many common dealers. Jaccard's Coefficient(JC) is a widely used similarity metric that quantifies this intuition where the number of common neighbours between two vertices is normalised by the total number of neighbors they have.

$$JC(uv) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|} \tag{2}$$

  Note that, in Equation 2, for any dealer x, $\Gamma(x)$ refers to the set of all dealers who have trade transactions (sales or purchases) with x.
- Flow Ratio: Assume that a vertex(dealer) o lead to the formation of a potential edge $\overrightarrow{uv}$, *i.e*., o is the out-neighbor of vertex v and the in-neighbor of vertex u. In this setting, we observe that triads generally tend to have a huge number of high-cash transactions from vertex v to u *via*. vertex o, as given in Figure 3, as opposed to the ones that does not form the triad. Keeping that in mind we define the variable Flow Ratio(FR).

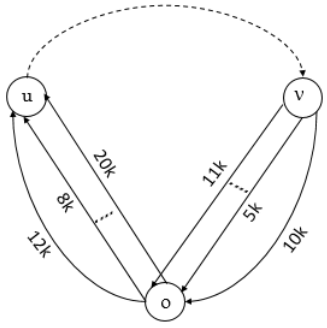

Figure 3: Triad with a large number of high-cash transactions from v to u *via*. o

$$FR(\overrightarrow{vou}) = \frac{Flow\_Count(\overrightarrow{vou}) * Flow\_Amount(\overrightarrow{vou})}{Flow\_Count(\overrightarrow{vou}) + Flow\_Amount(\overrightarrow{vou})} \tag{3}$$

  where, $Flow\_Count(\overrightarrow{vou})$ is the minimum value among the total number of edges from v to o and the total number of edges from o to u, and, $Flow\_Amount(\overrightarrow{vou})$ is the minimum value among the total weight of edges from v to o and the total weight of edges from o to u.
- Two Cycles: Recall from Section 1 that almost all circular trading cases are committed by the collusion of two or three dealers, which forms two-cycles or triads, respectively. Naturally, the dealers who are involved in the formation of triads

may already be performing heavy circular trade with other dealers involving two-cycles. It is worthy to note that, in many cases, the triad itself is formed due to a string of two-cycles as explained in Figure 4. Following this observation, we define the Two Cycles(TC) parameter.

$$TC(x) = |in - neigh(x) \cap out - neigh(x)| \tag{4}$$

where, $in{-}neigh(x)$ and $out{-}neigh(x)$ are the in-neighborhood and the out-neighborhood of a vertex x, respectively. Here, for a potential edge $\overrightarrow{uv}$, we used the feature variable $TC(u)$, *i.e.*, the number of dealers forming two-cycles with the seller in a potential edge. We note that $TC(v)$ and other interaction variables derived from the Two Cycles parameter did not improve the accuracy of the model.

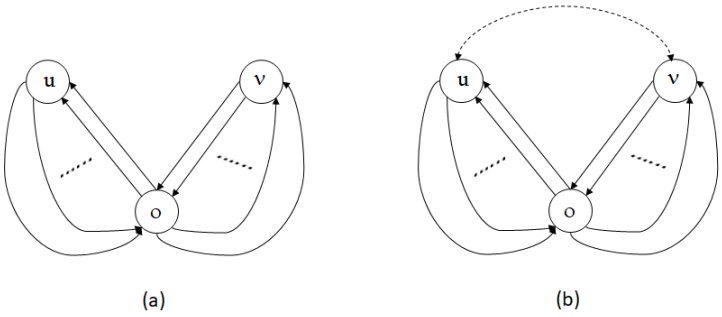

Figure 4: In (a) there are multiple two-cycles between v and o, and, u and o. In (b) addition of $uv^{->}$ or $vu^{->}$ forms triads $uvo^{->}$ or $vuo^{->}$, respectively.

- Interaction Variable: The Interaction Variable(IV) between the Jaccard's Coefficient and the Flow Ratio is the fourth variable used in the model.

$$IV(\overrightarrow{uv}) = JC(uv) \times FR(\overrightarrow{vou}) \tag{5}$$

Intuitively, this variable means that when two dealers share a high density of mutual neighbors and has a large flow of high cost transactions among them through another dealer, then, there is a chance for a direct link to form between them that aids the formation of a triad.

- Personalised PageRank: PageRank algorithm [4] is widely famous as the algorithm used by Google's search engine to rank web-pages in their search results. The key objective of the PageRank algorithm is to determine the importance of a web-page by taking into account the importance of the web-pages that are hyperlinked to it. In a graph-theoretic perspective, given the graph $G = (V, E)$, PageRank measures the stationary probability distribution of a custom random walk, as given below, starting from a random vertex in V and continues this walk until the PageRank vector converges.
  - At each iteration the walker either *jumps* to a random vertex with a predefined probability $(1 - p)$, or,
  - with probability p the walker $follows$ a randomly chosen outgoing edge of the current vertex.

Here we have used a personalised version [22] of the PageRank which is same as the PageRank algorithm other than the fact that *jumps* are always made to a particular vertex. In fact, as given in Equation 8, with probability $(1 - p)$ the random walker *jumps* to vertex u of the potential edge $\overrightarrow{uv}$. The motivation behind this technique is explained below.

Input to the personalised PageRank technique mentioned in Equation 8 is not the original graph G but a transformed version of G where edges that are part of a two-cycle are only being used. The *Two Cycles* parameter, defined earlier, makes use of the number of vertices making two cycles with the vertices of the potential edge. The personalised PageRank variable generalises the *Two Cycles* variable in a more efficient way by going into deeper levels of the network as explained below.

Even though *Two Cycles* is a good indicator for two dealers being involved in future triads, some dealers can be careful enough not to directly get involved in two-cycles but they may be involved with other dealers involved in performing circular trade using two-cycles. This trend where *'birds of the same feather flock together'* is expected among fraudulent dealers since it helps them to organise the crime in a more efficient and secure way. Motivated with this idea we redefine the edges of graph G using a normalised function of the number of two-cycles present between two vertices in G. The same is given in Equation 6. Here, if two vertices v_i and v_j forms two-cycles in graph G, then the minimum number of edges among v_i to v_j and v_j to v_i is assigned (after normalisation) as the weight of the directed edge (v_i, v_j) in the new graph (G').

Let A be an $n \times n$ matrix, where $n = |V|$, whose (i, j)-entry is defined as follows:

$$a_{ij} = \begin{cases} cycle(v_i, v_j), & \text{if a two-cycle exists between } v_i \text{ and } v_j \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where,

$$cycle(v_i, v_j) = \frac{min\Big(count(v_i, v_j), count(v_j, v_i)\Big)}{\sum_{i=1}^{n} min\Big(count(v_i, v_j), count(v_j, v_i)\Big)} \tag{7}$$

where, $count(v_i, v_j)$ gives the total number of edges (multiple edges) in $\overrightarrow{E}$ from v_i to v_j.

Let G' be the weighted directed graph represented by the adjacency matrix A in Equation 6. For the sake of simplicity, assume that G' is connected. In this scenario, one can easily observe that the matrix A is asymmetric and column-stochastic (column-normalised). Then the PageRank vector or the steady-state probability vector $\overrightarrow{\pi_u}$ of graph G' satisfies

$$\overrightarrow{\pi_u} = pA\overrightarrow{\pi_u} + (1 - p)\overrightarrow{q_u} \tag{8}$$

where p is the damping factor, generally set to 0.85, and $\overrightarrow{q_u}$ is an $(n \times 1)$ vector with 1 in the u-th row and and 0 otherwise. Note that vector $\overrightarrow{q_u}$ is used for personalisation. The random walk is performed as following:

Start the random walk by initialising the PageRank vector $\overrightarrow{\pi_u}$ to all ones. During each walk, update $\overrightarrow{\pi_u}$ to the new PageRank vector obtained after running Equation 8. Repeatedly perform the random walk until the PageRank vector converges, *i.e.*, $|\Delta\overrightarrow{\pi_u}|$ becomes negligible. It will converge since matrix A is column-normalised and the proof for the same is given in [21]. Now the PageRank vector $\overrightarrow{\pi_u}$ will be containing a list of ranking values to the vertices in G' with respect to vertex u. The higher the value of the v^{th} entry of the PageRank vector ($\overrightarrow{\pi_u}(v)$), the more probable it is for the potential edge $\overrightarrow{uv}$ to be formed in the future.

For classification, we built four models, *viz.*, KNN (K-Nearest Neighbors), LR (Logistic Regression), RF (Random Forest) and SVM (Support Vector Machine). As shown in Section 4, LR worked the best among all the four different classifiers. In the following Section, we give a brief overview of the LR classifier.

3.3 Logistic Regression

We built the logistic regression classifier combining the evidences derived from five feature variables to predict whether a *potential edge* will form in the future. Logistic regression learns a model of the form

$$P(1|x) = \frac{1}{1 + e^{-(c_0 + \sum_{i=1}^{n} c_i x_i)}} \tag{9}$$

where x represents the vector containing the feature variables $(x_1, \ldots, x_n)$ and $c_0, \ldots, c_n$ are the coefficients estimated from the training data.

4 EXPERIMENTAL RESULTS

4.1 Model Selection

Figure 5 shows the correlation among the four feature variables.

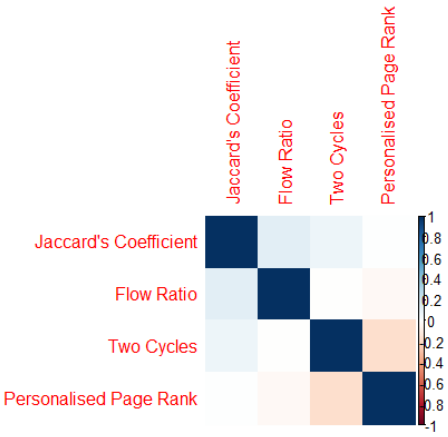

Figure 5: Feature correlation plot

As mentioned before, the LR model performs the best among all the four classification models. In LR, we have altered the cutoff value according to the miss-classification costs. For the taxation officials, miss-classifying fraudsters is more costly than miss-classifying genuine dealers [18].

They are built using the statistical programming language **R**[1]. Plot given in Figure 6 compares the performance of the four different models. Figure 7 shows the distribution of models' accuracy as density plots. Figure 8 shows the scatter plot matrix of all the fold results of a model against the same fold results for all the other models. Pairwise scatter plots of all the models are compared. Observing the scatter plots, LR and SVM models look highly correlated as does LR and KNN. KNN and RF are weakly correlated. It can be observed from these plots that the logistic regression model gives the best performance.

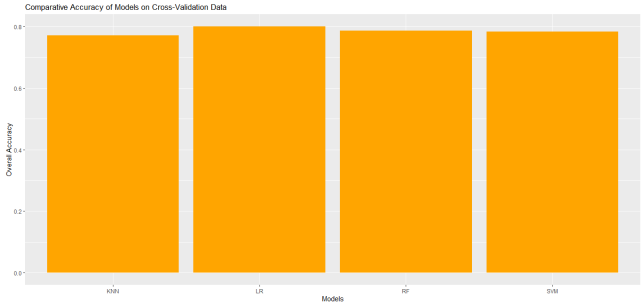

Figure 6: Model comparison plot

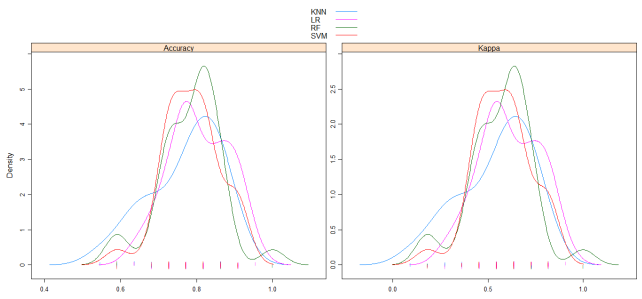

Figure 7: Density plot

4.2 Model Parametric Coefficients

Figure 9 shows the parametric coefficients of the logistic regression model. As one can observe, the p-value of the personalised PageRank parameter is the lowest among all and hence it becomes the most relevant feature variable in this model.

The relation between log of odds of dependent variable Link and independent variables Jaccard's Coefficient, Flow Ratio, Two Cycles and Personalised PageRank as shown in Figures 10, 11, 12 and 13, respectively, are linear.

[1] https://www.r-project.org/

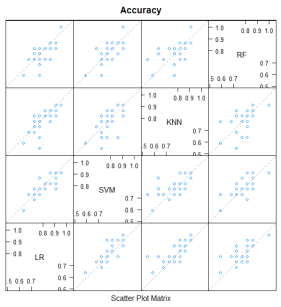

Figure 8: Scatter plot

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-3.53036	0.52673	-6.702	2.05e-11
Jaccard's Coefficient	16.58327	4.14317	4.003	6.27e-05
Flow Ratio	0.07145	0.01670	4.279	1.87e-05
Two Cycles	17.76736	5.83902	3.043	0.002343
Interaction Variable	-0.41836	0.12173	-3.437	0.000589
Personalised PageRank	0.27286	0.05665	4.817	1.46e-06

Figure 9: Parametric Coefficients

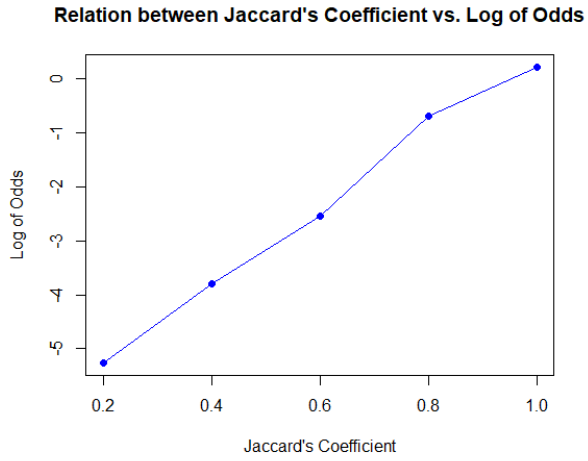

Figure 10: Jaccard's Coefficient Vs Log of Odds

4.3 Model Performance

In order to evaluate the performance of this model, we compute accuracy, precision, recall and F1-score. We computed all these parameters for train as well as for test dataset as shown in Table 5. One can observe that the accuracy of the test dataset is 80%.

We validated the model using the following measures:

- Concordance Measure: Concordance value is 0.85671 and discordance value is 0.14329.
- ROC Curve: As shown in Figure 14, the area under the train dataset ROC curve is 0.894 and the test dataset ROC curve is 0.856. Since there is not much difference between the ROC curves, one can conclude that the model is not over fitting. Since the area under the train dataset ROC curve is more than 0.70, one can say that model is not under fitting.
- Lift chart: Lift chart measures the effectiveness of a predictive model and it is the ratio between the results obtained with and without the predictive model. Figure 15 depicts the lift chart for the model.

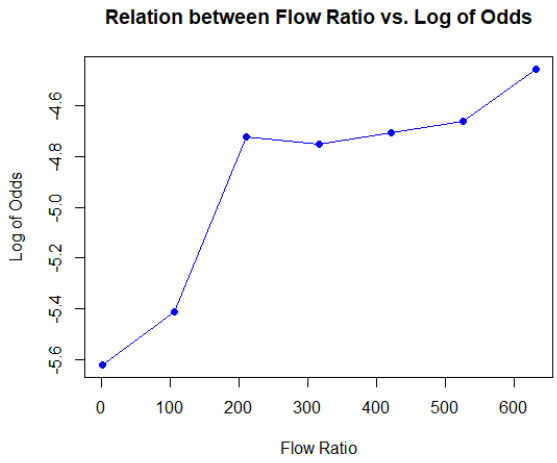

Figure 11: Flow Ratio Vs Log of Odds

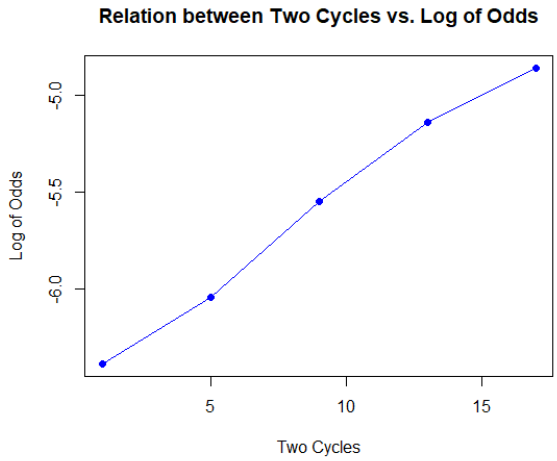

Figure 12: Two Cycles Vs Log of Odds

Table 5: Performance evaluation of the model

Dataset	Accuracy	Precision	Recall	F1-Score
Train	0.8090	0.85416	0.7454	0.7960
Test	0.80	0.8383	0.7410	0.7866

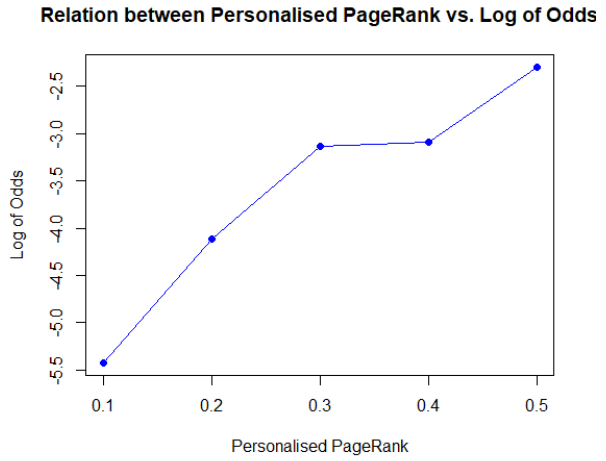

Figure 13: Personalised PageRank Vs Log of Odds

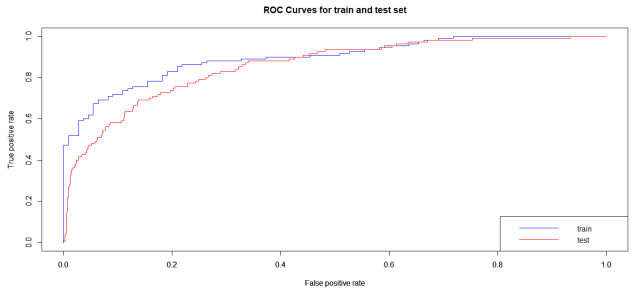

Figure 14: ROC curves for train and test data

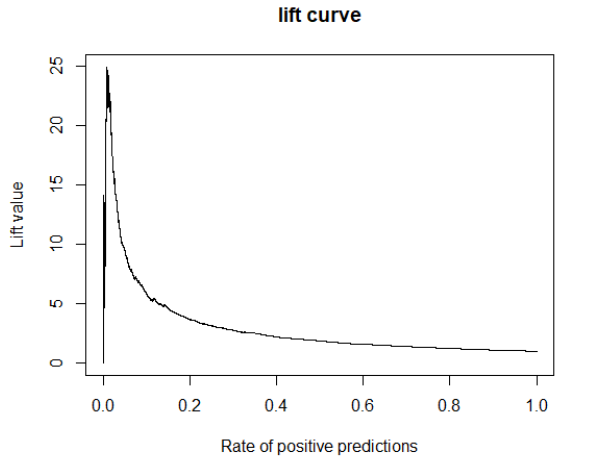

Figure 15: Lift Chart

5 CONCLUSION

In this paper, we devised four classification models for predicting three-cycles in a trade-network. These three-cycles are undesired patterns in the network as they are used by dealers to advance the illegal practise of circular trading. Circular trading is practiced by tax evaders by issuing sales invoices among a closed group without any value-addition and without any actual supply of goods among them. It helps to hype their turnover, and this in-turn allows them to perform a multiple varieties of other more critical financial crimes. In this work, we entwine domain specific observations with link prediction parameters to build the binary classifiers. We observed that the logistic regression model performed the best among all four classifiers and it predicted with 80% accuracy whether a three-cycle would be formed in the near future or not. Predicting three cycles before they happen significantly helped the tax enforcement officers to increase the tax revenue of the state.

Here, to create the feature variables for the classification models, we have extensively exploited the presence of two-cycles in the trade network. From Table 2 it can be observed that two-cycles comprises the majority of circular trading cases since it is easier to perform. However, we have not found any efficient way to predict them. Trying to model the formation of two-cycles is a worthy endeavor for the future.

ACKNOWLEDGMENTS

We are very grateful to the Telangana state government, India, for sharing the commercial tax dataset, which is used in this work. This work has been supported by Visvesvaraya PhD Scheme for Electronics and IT, Media Lab Asia, grant number EE/2015-16/023/MLB/MZAK/0176.

REFERENCES

[1] Lada Adamic and Eytan Adar. 2003. Friends and Neighbors on the Web. *Social Networks* 25 (07 2003), 211–230. https://doi.org/10.1016/S0378-8733(03)00009-1

[2] Zhenisbek Assylbekov, Igor Melnykov, Rustam Bekishev, Assel Baltabayeva, Dariya Bissengaliyeva, and Eldar Mamlin. 2016. *Detecting Value-Added Tax Evasion by Business Entities of Kazakhstan.* 37–49. https://doi.org/10.1007/978-3-319-39630-9_4

[3] Pietro A. Bianchi and Miguel Minutti-Meza. 2016. Professional Networks and Client Tax Avoidance: Evidence from the Italian Statutory Audit Regime. *SSRN Electronic Journal* (01 2016). https://doi.org/10.2139/ssrn.2601570

[4] Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30, 1–7 (1998), 107–117. http://citeseer.ist.psu.edu/brin98anatomy.html

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly Detection: A Survey. *ACM Comput. Surv.* 41 (07 2009). https://doi.org/10.1145/1541880.1541882

[6] Godbole Committee. 1998. *Report on Economic Reforms of Jammu and Kashmir.* Technical Report. Ministry of Finance, Government of Jammu and Kashmir.

[7] Fintax Consultancy. 2019. *All About Circular Trading under GST.* Retrieved May 15, 2019 from https://fintaxconsultancy.co.in/Blog/2019/05/15/all-about-circular-trading-under-gst/

[8] Sachin Dave. 2019. *Circular trading & GST evasion charges: Taxman may have to review arrest strategy.* Retrieved April 20, 2019 from https://economictimes.indiatimes.com/news/economy/policy/circular-trading-gst-evasion-charges-taxman-may-have-to-review-arrest-strategy/articleshow/68961669.cms

[9] Pamela González and Juan Velasquez. 2013. Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications* 40 (04 2013), 1427–1436. https://doi.org/10.1016/j.eswa.2012.08.051

[10] Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. 2004. Combating Web Spam with TrustRank. *Intl Conference on Very Large Data Bases*, 576–587. https://doi.org/10.1016/B978-012088469-8/50052-8

[11] Hussein Issa and Miklos Vasarhelyi. 2011. Application of Anomaly Detection Techniques to Identify Fraudulent Refunds. *SSRN Electronic Journal* (08 2011).

https://doi.org/10.2139/ssrn.1910468

[12] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Predicting Positive and Negative Links in Online Social Networks. *Proceedings of the 19th International Conference on World Wide Web, WWW '10* (03 2010). https://doi.org/10.1145/1772690.1772756

[13] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Signed Networks in Social Media. *Conference on Human Factors in Computing Systems - Proceedings* 2 (03 2010). https://doi.org/10.1145/1753326.1753532

[14] Víctor Martínez, Fernando Berzal, and Juan-Carlos Cubero. 2016. A Survey of Link Prediction in Complex Networks. *Comput. Surveys* 49 (12 2016), 69. https://doi.org/10.1145/3012704

[15] Jithin Mathews, Priya Mehta, Suryamukhi K., Dikshant Bisht, Sobhan Chintapalli, and S.V. Rao. 2018. Regression Analysis towards Estimating Tax Evasion in Goods and Services Tax. 758–761. https://doi.org/10.1109/WI.2018.00011

[16] Girish Palshikar and Manoj Apte. 2008. Collusion set detection using graph clustering. *Data Min. Knowl. Discov.* 16 (04 2008), 135–164. https://doi.org/10.1007/s10618-007-0076-8

[17] Mehdi Rad and Asadollah Shahbahrami. 2015. High performance implementation of tax fraud detection algorithm. 6–9. https://doi.org/10.1109/SPIS.2015.7422302

[18] P. Ravisankar, V. Ravi, G. Raghava Rao, and I. Bose. 2011. Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems* 50, 2 (2011), 491–500.

[19] Daniel Roux, Boris Pérez Gutiérrez, Andres Moreno, Pilar Villamil, and César Figueroa. 2018. Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach. 215–222. https://doi.org/10.1145/3219819.3219878

[20] Yusuf Sahin and Ekrem Duman. 2011. Detecting credit card fraud by ANN and logistic regression. *INISTA 2011 - 2011 International Symposium on INnovations in Intelligent SysTems and Applications* (06 2011). https://doi.org/10.1109/INISTA.2011.5946108

[21] Gilbert Strang. 2003. *Introduction to Linear Algebra.*

[22] J. Sun, Huiming Qu, Deepayan Chakrabarti, and C. Faloutsos. 2005. Neighborhood formation and anomaly detection in bipartite graph. *Proceedings of IEEE International Conference on Data Mining*, 8 pp.–. https://doi.org/10.1109/ICDM.2005.103

[23] Junjie Wang, Shuigeng Zhou, and Jihong Guan. 2011. Detecting Collusive Cliques in Futures Markets Based on Trading Behaviors from Real Data. *Neurocomputing* 92 (10 2011). https://doi.org/10.1016/j.neucom.2011.11.022

[24] Shuhan Yuan, Xintao Wu, Jun Li, and Aidong Lu. 2017. Spectrum-based Deep Neural Networks for Fraud Detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*. ACM, 2419–2422. https://doi.org/10.1145/3132847.3133139

[25] Muhan Zhang and Yixin Chen. 2018. Link Prediction Based on Graph Neural Networks. (02 2018).