Journal of Engineering Sciences
Vol 11, Issue 11, NOV/2020
ISSN NO:0377-9254

Codetect: Financial Fraud Detection Using Anamaly Feature Detection

Dr.A.Anjaiah [1](Associate professor)
Department of Computer Science and Engineering
St. Peter's Engineering College, Dullapally, Maisammaguda,
Medchal, Hyderabad, Telangana 500043
anjaiah@stpetershyd.com[1]

P.Haripriya[2]
(UG Student)
Department of Computer Science and Engineering
St. Peter's Engineering College, Dullapally, Maisammaguda, Medchal, Hyderabad, Telangana 500043

I.GokulRaj[3]
(UG Student)
Department of Computer Science and Engineering
St. Peter's Engineering College, Dullapally, Maisammaguda, Medchal, Hyderabad, Telangana 500043

R.Sunilkumar[4]
(UG Student)
Department of Computer Science and Engineering
St. Peter's Engineering College, Dullapally, Maisammaguda, Medchal, Hyderabad, Telangana 500043

Abstract: ***Financial fraud, such as money laundering, is known to be a serious process of crime that makes illegitimately obtained funds go to terrorism or other criminal activity. This kind of illegal activities involve complex networks of trade and financial transactions, which makes it difficult to detect the fraud entities and discover the features of fraud. Fortunately, trading/transaction network and features of entities in the network can be constructed from the complex networks of the trade and financial transactions. The trading/transaction network reveals the interaction between entities, and thus anomaly detection on trading networks can reveal the entities involved in the fraud activity; while features of entities are the description of entities, and anomaly detection on features can reeffect details of the fraud activities. Thus, network and features provide complementary information for fraud detection, which has potential to improve fraud detection performance. However, the majority of existing methods focus on networks or features information separately, which doesnot utilize both information. In this paper, we propose a novel fraud detection framework, CoDetect, which can leverage both network information and feature information for financial fraud detection. In addition, the CoDetect can simultaneously detecting financial fraud activities and the feature patterns associated with the fraud activities. Extensive experiments on both synthetic data and real-world data demonstrate the efficiency and the effectiveness of the proposed framework in combating financial fraud, especially for money laundering.***

Keywords - Financial fraud, Anamaly detection, Credit card fraud.

I.INTRODUCTION

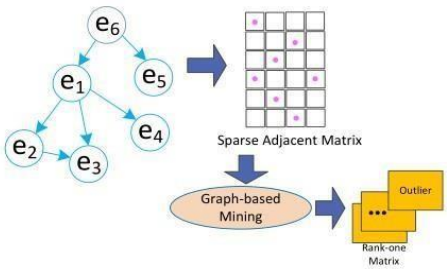
In recent years, financial fraud activities such as credit card fraud, money laundering, increase gradually. These activities cause the loss of personal and/or enterprises' properties. Even worse, they endanger the security of nation because the protfrom fraud may go to terrorism [1], [25]. Thus, accurately detecting financial fraud and tracing fraud are necessary and urgent. However, financial fraud detection is not an easy task due to the complex trading networks and transactions involved. Taking money laundering as an example, money laundering is denied as the process of using trades to move money/goods with the intent of obscuring the true origin of funds. Usually, the prices, quantity or quality of goods on an invoice of money laundering are fake purposely. The misrepresentation of prices, quantity or quality of goods on an invoice merely exposes slight difference from regular basis if we use these numbers as features to generate detection policy. Under certain circumstances, this kind of detector may work well with relatively stable trading entities. Unfortunately, the real world situation is more complicated, especially within Free Trade Zones (FTZs) where international trade involves complex procedures and exchange of information between trading entities. The fraud activities, especial money laundering, are deeper stealth. Money laundering activities may take different forms [1] such as the concealing

transportation of cash using trading operations; the acquisition and sale of intangibles; and related party transactions.

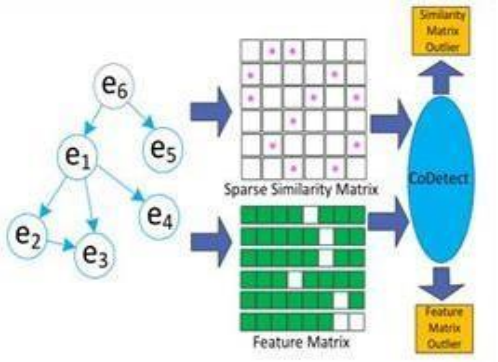
In contrast with other fraud activities, money laundering demonstrates special characteristic which presents high risk to financial system with obscuring the money trail, collectivization behavior and wild trading regions in FTZs.

Many fraud detection models work with attribute value data points that are generated from transactions data. Some aggregation methods are also used to enrich the information of data [28]. After generating feature points from transactions, supervised and unsupervised methods can be used to perform detection [26], [27], [34]. Usually, these data points are assumed to be independent and identically distributed (i.i.d.). However, the characteristic of money laundering is different from attribute-value data. The collectivization behavior means the data is inherently linked or partly linked. Obviously, trading activity involves at least two business entities. Linked data is patently not independent and identically distributed, which contradicts the assumptions of traditional supervised and unsupervised methods. On the other side, some linked data is auto correlated. For example, trading between business entity A and B implies that feature points A and B are correlated. Some features used to describe the properties of trading goods can be identical between A and B.

This characteristic of auto correlation reduce the effective size of data for learning. Furthermore, feature points don't fuse the interaction information in data. The relations between any business entities indicate the potential causality that means, if businesses on going, fraud entity can be located by other identfied fraud entity. This means the entity, which have connection with fraud entity, are suspicious. Consequently, feature based detection models with supervised or unsupervised methods have inherent limitation of incapacity of identifying what the fraud relations are. what the fraud relations are. Graph-based mining methods are one of the most important theories that attempt to identify relations between data points [3], [7], [13], as Fig. 1(a) shows. Financial activities can be modeled as a directed graph, then a sparse adjacent matrix can represent this graph. With graph-mining method, the sparse matrix can be approximated as summation of low-rank matrix and outlier matrix. The outlier matrix is a sign of suspicious fraud activities. Exploiting the graph based mining provides a new perspective for fraud detection and enables us to do advanced research on fraud detection. With the fraud activities detected by graph-based detection technique we are able to draw the conclusion that several business entities involved in fraud, however, we still don't know how these fraud activities are operated and why these activities labeled as fraud, i.e., the detailed features of the fraud activities. The majority of this how-and-why information is fused in features points, which have essential meaning for financial fraud detection because of the tracing necessity.



(a).Existing fraud detection framework



(b).The Proposed framework

Fig 1:System Overview

II. RESEARCH METHOD

A).Graph Matrix and Feature Matrix from STDLAT

In this subsection, we present how we build graph matrix and feature matrix from SDLAT information. Since SDLAT contains the source organizationand destination organization, we can develop a system $G = \{v, \}$ where$V = v_1, \ldots, v_N$ is a lot of N hubs with every hub being an organization and $\varepsilon \subset v \times v$ is a set of edges. In the event that is t he source organization and v_j is the goal organization, we include and edge e_{ij} $= 1$ between v_i and v_j. Notwithstanding the system,every hub is additionally connected with a set of properties, for example,location, asset, tax status. We exploits $F \in R^{N\times d}$ to disclose the feature-matrix, where d is the component of the features. The system G contains the associations among organizations. From the system structure of G, we may ready to recognize scenario 1 and 3 financial fraud. $e_{ij} = 1$ just implies that there's an interaction fromv_i to . To mirror the resemblance between the source and goal organization, this doesn't mirror the cost of the goods or different assets and subsequently can't be utilized to identify scenario 2. To consolidate data for identifying scenario 2, we also exploits S_{ij} to disclose the weight between v_i and v_j . The weight S_{ij} is defined as:

$$S = e^{\frac{f_i - f_j}{\sigma^2}} \quad (1)$$

where f_i implies the I-th column of **F** and **σ** is a scalar to control the size of the weight. Along these lines, **S** is the weighted graph data and **F** is the feature-matrix. The issue is officially characterized as:

Given graph matrix **S** and feature matrix **F**, discover a function f which can at the same time identify fraud activities and trace the properties of the fraud.

B). Anamaly Detection on Graph Matrix

In genuine world, the exchange are ordinarily among organizations of comparable kind, i.e., organizations that manage comparable business are bound to have interaction. For instance, for an IT organization v_i it is bound to see v_j to have exchange/business with IT organizations than natural product organizations. This reality makes the graph-matrix which comprises of block-structures.

Organizations which are within a similar block are of comparative business type and there are a larger number of collaborations of organizations inside each block than that of between blocks. Alternatively, the graph-matrix is low- rank [4]. Thus, we can provide as **S**:

$$S_{ij} = UV^i{}_s + R_s \quad (2)$$

organizations of a similar kind, for example, the value, the where$U \in R^{N\times r}$ and $V_S \in R^{N\times r}$ are two low-rank latent feature matrix with $r << N$.The interface between v_i and v_j is recouped by the connection between the latent highlights of u_i and v_j as $U\ i, : V_S\ j, :^T, UV^T$ will

give a low rank-matrix, which fundamentally recuperates the inner blocks associations. R_S is the residual-matrix, which principally incorporates the association between the blocks. As we probably know, the fraud-transaction is uncommon, every two organizations in exchanging is reliant. In graph- mining, low-rank matrix is utilized to give the transaction information[4]. Since the connections between blocks, i.e., the exchange between organizations of various kinds, are uncommon and suspicious, , R_S can be utilized to catch the suspicious communication and would thus be able to be utilized to detect fraud [7]. Given the way that most of collaborations are typical and are not financial-fraud, we would anticipate that the caught financial-fraud should be scanty. In light of this, we include the l_1 norm , in order to make □□ inadequate and can catch genuine financial-fraud. At that point the objective function accompanied as:

$$\min_{U,V_S,R_S} \|R_s\|_1 \quad (3)$$

$$s.t. = V^T{}_S + R_S$$

Since **U** is the latent highlights of organizations and organizations structure gatherings, i.e., a few organizations do comparative business, we would expect the latent-features of organizations within a same-group have comparative latent-features. In view of this, we include the orthogonal constraint **U**, which is generally utilized for separating group of features [8]. After summing-up the orthogonal constraint, Equation (3) becomes:

$$\min_{U,V_S,R_S} \|R_s\|_1$$

$$s.t.\ S = UV^T{}_S + R_S$$

$$\mathbf{U}^T\mathbf{U} = \mathbf{I} \quad (4)$$

Here, norm 1 is exploited to ensure the detected fraud is rare. **U** is pseudo class label.

A) Anomaly Detection on Feature Matrix

With residual matrix **R** we can without much of a stretch explain what number of business elements include in fraud and what is the pattern for the fraud, for example merge (or) ring. There are yet tremendous of data we don't know about the fraud, for example, position, value, tax and so on which can be written by SDLAT include. Those fraud data is important to financial related official for fraud tracking. In this way, anomaly-detection on matrix **F** is essential. Concerning typical finance related business, we would expect comparative component examples to have within position. Hence, the feature-matrix **F** is normally low-

rank as organizations of a similar kind has comparative feature-

patterns [25]. In light of this perception, we initially deteriorate the feature-matrix as **F** as:

$$\mathbf{F} = \mathbf{U}\mathbf{V}^T + \mathbf{R}_f \quad (5)$$

where $U \in R^{N\times r}$ is the latent representations of the companies and V_f are the latent representations of the SDLAT features. $\mathbf{R}_f$ is the residual matrix. For features that can't be all around reproduced, the relating residual will be huge, which mirrors the anomaly highlights. Along these lines, with the residual matrix , we can track the fraud- patterns. Since the most of the organizations don't include in financial-fraud, we can expect that the residual matrix $\mathbf{R}_f$ is scanty. Along these

$$s.t.\ S = UV^T + R_f \quad f$$

$$\mathbf{U}^T\mathbf{U} = 1 \ (7)$$

D) The CoDetect Framework

Equation (4) reflects the graph-matrix **S** to detect fraud actions since (7) manages **F** to track the fraud-patterns. To intact use these two matrices for at the same time detecting financial-fraud and tracing fraud-patterns, we can articulate
(4) and (7), which accompanies in theobjective-function of CoDetect:

$$\arg\min_{U} || R_s ||_1 + \alpha || R_f ||_1$$

$$s.t.\ S = U * V^T + R_s \quad s$$

$$F = U * V^T_f + R_f \quad (8)$$

Where α is a scalar to influence the contribution of the graph matrix **S** and feature matrix **F**. The latent organization feature-matrix **U** is found out from both **S** and **F** as by the
requirement $s.t.S = U*V^T_S + R_S$ and $F = U * V^T_f + R_f$ Thus

data of **S** and **F** can move through **U** and in this way proposed CoDetect is a bounded structure that exploits both **U** and **V** all the while.

We assess the detection precision on similar matrix and feature-matrix separately. We infuse three fraud designs into two dataset separately. We initially accomplish the examinations by CoDetect, Robust PCA and SVD for the correlation of accuracy on likeness. RPCA and SVD are utilized to analyze top k rank components, at that point we acquire the residual-matrix by unique matrix less top k rank segments. Here k is set to 5. We preclude the parameter investigation and just report the best execution on RPCA and SVD. We rehash the analyses multiple times

lines , we likewise include l_1 standard
$\mathbf{R}_f$ to make it meager, which gives us the objective-function
as:

$$\min_{U,,R_f} || R_f ||_1 \quad f \quad (6)$$

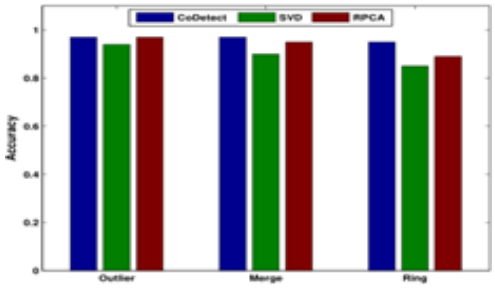
$$s.t.\ S = UV^T + R_f \quad f$$

correspondingly, we include the orthogonal constraint on **U**

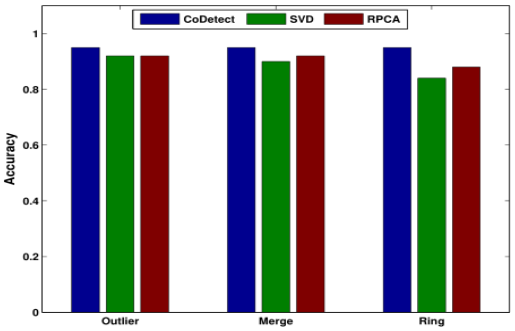
to formulate it discriminative as:

$$\min_{U,,R_f} || R_f ||_1$$

and report the mean exactness on similarity-matrix. From Fig. 2 we see that CoDetect and RPCA accomplishes high detection precision on similarity -matrix from synthetic-information and genuine information. We play out the examinations on feature-Matrix accuracy on all fraud patterns.



a) Similarity Matrix(synthetic data)



b) Similarity Matrix(real life data)

Fig 2: Detection accuracy on graph-based similarity matrix.

Time Performance Analysis. We assess the time performance here.The investigations are altogether performed on machine with Intel(R) Core(TM) i7 CUP @ 2.60GHz and 32GB memory, running Windows 7. Each test is rehashed multiple times and we report the meantime in second. We initially analyze the adaptability of CoDetect by retune the size of graph. We tune the size of graph from 5,000 t0 25,000 and tune the edge-number from 5 × 105 to 15 × 105, at that point infuse three fraud designs into each diagram . . At that point we assess the detection-time execution in term of second.

III. RESULT ANALYSIS

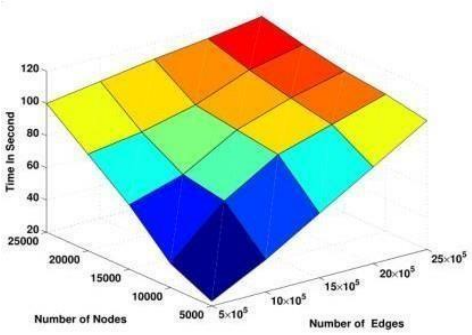


Fig 3: Detection time in second with different number of nodes and edges

We find that CoDetect unite to edge in 10 cycles for the most part. So we set the cycle to 10 so as to hinder the calculation cost. The outcome is introduced in Fig. 2. It tends to be seen that CoDetect scales directly with retune the graph size and number of edge. All the detection can be finished in limited-time. The following examinations are performed utilizing Iknow.com dataset with around 27,000 hubs and 5,600,000 edges. We think about the time execution of CoDetect, RPCA and SVD with various number of rank, r for registering the leftover network. The outcome is displayed in Fig. 3. Plainly, CoDetect accomplishes high time execution.

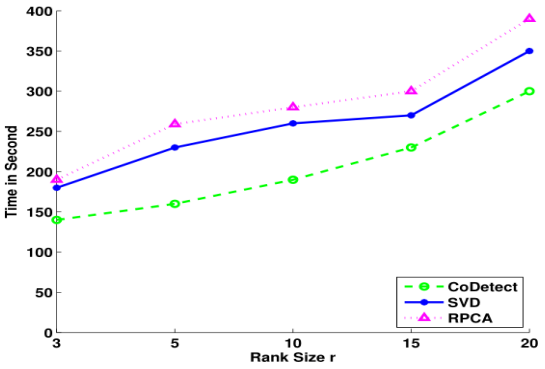


Fig 4: Comparision of time with different rank size

IV.CONCLUSION

We propose another system, CoDetect, which can perform fraud-detection on graph based similarity-matrix and feature-matrix at the same time. It acquaints another path with uncover the idea of budgetary activities from fraud examples to suspicious property. Besides, the system gives an increasingly interpretable approach to recognize the fraud on sparse-matrix. Test results on manufactured and certifiable informational collections demonstrate that the proposed structure (CoDetect) can viably distinguish the fraud designs just as suspicious highlights. With this co- detection system, administrators in financial supervision can detect the fraud designs as well as trace the origin of fraud with apprehensive feature.

V. ACKNOWLEDGEMENT

This research paper has been equipped as a result of discussion among the faculty members of my college. The authors like to thank Dr..Anjaiah Adepu (Associate Professor) of St.Peter's Engineering College, Hyderebad, Telangana for their constant guidance and support during the research work time period.

VI. REFERENCES

[1] C. Sullivan and E. Smith. ``Trade-Based Money Laundering: Risks and Regulatory Responses," Social Sci. Electron. Publishing, 2012, p. 6.

[2] United Press International. (May 2009). *Trade-Based Money Laundering Flourishing*.[Online].Available:http://www.upi.com/TopNews/2009/05/11/Trade-based-money-laundering-_ourishing/UPI-17331242061466

[3] L. Akoglu, M. McGlohon, and C. Faloutsos, ``Odd Ball: Spotting anomalies in weighted graphs," in *Proc. Pacic-Asia Conf. Knowl. Discovery Data Mining*, 2010, pp. 410_421.

[4] V. Chandola, A. Banerjee, and V. Kumar, ``Anomaly detection: A survey,"*ACM Comput. Surv.*, vol. 41, no. 3, 2009, Art. no. 15.

[5] W. Eberle and L. Holder, ``Mining for structural anomalies in graph-baseddata," in *Proc. DMin*, 2007, pp. 376_389.

[6] C. C. Noble and D. J. Cook, ``Graph-based anomaly detection," in *Proc.9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2003,pp. 631_636.

[7] H. Tong and C.-Y. Lin, ``Non-negative residual matrix factorization with application to graph anomaly detection," in *Proc. SIAM Int. Conf. Data Mining*, 2011, pp. 1_11.

[8] S.Wang, J. Tang, and H. Liu, ``Embedded unsupervised feature selection," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 470_476.

[9] Z. Lin, M. Chen, and Y. Ma. (2010). ``The Augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices." [Online].Available: https://arxiv.org/abs/1009.5055.

[10] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos, ``Neighborhood formationand anomaly detection in bipartite graphs," in *Proc. 15th IEEE Int. Conf.Data Mining*, Nov. 2005, p. 8.

[11] A. Patcha and J.-M. Park, ``An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Comput. Netw.*, vol. 51,no. 12, pp. 3448_3470, Aug. 2007.

[12] W. Li, V. Mahadevan, and N. Vasconcelos,``Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*,vol. 36, no. 1, p. 18_32, Jan. 2014.

[13] K. Henderson *et al.*, ``It's who you know: Graph mining using recursive structural features," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 663_671.

[14] F. Keller, E. Müller, and K. Bohm, ``HiCS: High contrast subspaces for density-based outlier ranking," in *Proc. ICDE*, Apr. 2012, pp. 1037_1048.

[15] D. Koutra, E. Papalexakis, and C. Faloutsos, ``Tensorsplat: Spotting latentanomalies in time," in *Proc. PCI*, Oct. 2012, pp. 144_149.