General

<table>
<tr><td>ID[1]</td><td colspan="4"></td></tr>
<tr><td>Use case name</td><td colspan="4">Detection of frauds based on collusions</td></tr>
<tr><td>Context</td><td colspan="4">Fintech</td></tr>
<tr><td>Application domain</td><td colspan="4">On-premise systems</td></tr>
<tr><td>Status</td><td colspan="4">In operation</td></tr>
<tr><td rowspan="2">Contributor</td><td>Name</td><td>Affiliation</td><td colspan="2">Contact</td></tr>
<tr><td>Girish Palshikar
C. Anantaram</td><td>Tata Consultancy Services Ltd.</td><td colspan="2">c.anantaram@tcs.com</td></tr>
<tr><td>Scope[2]</td><td colspan="4">Validating the predicted collusion set is effort-intensive and needs investigative and legal expertise</td></tr>
<tr><td>Objective(s)</td><td colspan="4">Automatic unsupervised detection of frauds based on collusions</td></tr>
<tr><td rowspan="2">Narrative</td><td>Short description (not more than 150 words)</td><td colspan="3">Our tool includes a set of unsupervised machine learning algorithms to detect collusion-based frauds, particularly, circular trading and price manipulation in stock market trading</td></tr>
<tr><td>Complete description</td><td colspan="3">Frauds are prevalent across all industries; and they are particularly severe in today’s computerized, web-connected, mobile-accessible, and cloud-enabled business environments. An FBI report states that the insurance industry in the US, which consists of over 7000 companies and collects over $1 trillion in premiums, loses about $40 billion annually in frauds in the non-health insurance sector alone. The aggregate size of the 52 regulated stock exchanges across the world (total market capitalization) was $55 trillion as on Dec. 2012. Given the money involved, it is not surprising that the stock market is a target of frauds.

Many malpractices in stock market trading, e.g. circular trading and price manipulation—use the modus operandi of collusion. Informally, a set of traders is a candidate collusion set when they have “heavy trading” among themselves, as compared to their trading with others. We formalize the problem of detection of collusion sets, if any, in a given trading database. We show that naïve approaches are inefficient for real-life situations. We adapt and apply two well-known graph clustering algorithms for this problem. We also propose a new graph clustering algorithm, specifically tailored for detecting collusion Sets; further, we establish a combined collusion set. Treating individual experiments as evidence, this approach allows us to quantify the confidence (or belief) in the candidate collusion sets. We have carried out detailed simulation experiments to demonstrate effectiveness of the proposed algorithms. The system is also operational in a government organization. Note that all our collusion detection algorithms are completely unsupervised and do not need any training data.</td></tr>
<tr><td rowspan="2">Key performance indicators (KPIs)</td><td>ID</td><td>Name</td><td>Description</td><td>Reference to mentioned use case objectives</td></tr>
<tr><td>1</td><td>Prediction accuracy</td><td>How many predicted collusion sets</td><td>Improve accuracy</td></tr>
</table>

			were actually involved in frauds	
AI features	Taks(s)	Knowledge processing & discovery		
	Method(s)[3]	Machine learning		
	Hardware[4]	Windows		
	Terms and concepts used[5]			
Challenges and issues	Challenges: Actual examples of collusion-based frauds may not be available easily, even for evaluation and testing			
Societal concerns				

Data (optional)

Data characteristics	
Description	
Source[6]	
Type[7]	
Volume (size)	
Velocity (e.g. real time)[8]	
Variety (multiple datasets)[9]	
Variability (rate of change)[10]	
Quality[11]	

Process scenario (optional)

Scenario conditions					
No.	Scenario name	Scenario description	Triggering event	Pre-condition[12]	Post-condition[13]

Training (optional)

Scenario name	Training				
Step No.	Event[14]	Name of process/Activity[15]	Primary actor	Description of process/activity	Requirement

Specification of training data[16]	

Evaluation (optional)

Scenario name	Evaluation				
Step No.	Event[17]	Name of process/Activity[18]	Primary actor	Description of process/activity	Requirement

Input of evaluation[19]	
Output of evaluation[20]	

Execution (optional)

Scenario name	Execution				
Step No.	Event[21]	Name of process/Activity[22]	Primary actor	Description of process/activity	Requirement

Input of Execution[23]	
Output of Execution[24]	

Retraining (optional)

Scenario name	Retraining				
Step No.	Event[25]	Name of process/Activity[26]	Primary actor	Description of process/activity	Requirement

Specification of retraining data[27]	

References

References						
No.	Type	Reference	Status	Impact on use case	Originator/organization	Link
1	Conference				Tata Consultancy Services Limited	D. K. Luna, G. K. Palshikar, M. Apte, A. Bhattacharya, *Finding Shell Company Accounts using Anomaly Detection*, **ACM India Joint International Conference on Data Science and Management (CoDS-COMAD 2018),** Goa, India, Jan 11-13, 2018
2	Journal				Tata Consultancy Services Limited	G. K. Palshikar, M. Apte, *Collusion Set Detection Using Graph Clustering*, vol. 16, no. 2, April 2008, **Data Mining and Knowledge Discovery** journal (Springer-Verlag), pp. 135 – 164
3	Book chapter				Tata Consultancy Services Limited	M. Apte, G.K. Palshikar, S. Baskaran, *Frauds in Online Social Networks: A Review*, accepted as a Book Chapter, in **Social Network and Surveillance for Society**, T. Ozyer and S. Bakshi (ed.s), to be published by Springer in 2018
4	Book chapter				Tata Consultancy Services Limited	G.K. Palshikar, M. Apte, *Financial Security against Money Laundering: A Survey*, Chapter 36 in B. Akhgar, H.R. Arabnia (Ed.s), **Emerging Trends in Information and Communication Technologies**

							Security, pp. 577 – 590, Elsevier (Morgan Kaufman), 2013

[1] Leave this cell blank.

[2] The scope defines the limits of the use case.

[3] AI method(s)/framework(s) used.

[4] Hardware system used.

[5] Terms and concepts listed here can be used to extend the work of WG 1 (AWI 22989 and AWI 23053) as necessary.

[6] Origin of data, which could be from instruments, IoT, web, surveys, commercial activity, or from simulations.

[7] Structured/unstructured Images, voices, text, gene sequences, and numerical. Composite: time-series, graph-structured

[8] The rate of flow at which the data is created, stored, analysed, or visualized.

[9] Data from a number of domains and a number of data types. The wider range of data formats, logical models, timescales, and semantics complicates the integration of the variety of data.

[10] Changes in data rate, format/structure, semantics, and/or quality.

[11] Completeness and accuracy of the data with respect to semantic content as well as syntactical of the data (such as presence of missing fields or incorrect values)

[12] Describe which condition(s) should have been met before this scenario happens.

[13] Describe which condition(s) should prevail after this scenario happens. The post-condition may also define "success" or "failure" conditions.

[14] The event that triggers the step. This might be completion of the previous event.

[15] Action verbs should be used when naming activity.

[16] Training data can be further specified.

[17] The event that triggers the step. This might be completion of the previous event.

[18] Action verbs should be used when naming activity.

[19] Specify input of evaluation.

[20] Specify output of evaluation.

[21] The event that triggers the step. This might be completion of the previous event.

[22] Action verbs should be used when naming activity.

[23] Specify input of evaluation.

[24] Specify output of evaluation.

[25] The event that triggers the step. This might be completion of the previous event.

[26] Action verbs should be used when naming activity.

[27] Retraining data can be further specified.