A Network Analysis Approach to Detect Collusive Fraud in Procurement Process

Abhinav Bajpai
Department of Data Science
Indian University – Bloomington
USA, 47405
abbajpai@iu.edu

Abstract

Collusive fraud occurs when organizations or individuals work together to circumvent legal processes and procedures. Often these relationships are hidden behind layers of transactions that are difficult to detect. Recent developments in network science allow us to take a fresh look at collusive fraud. In this study, we apply OddBall to detect outliers in a procurement dataset and also test whether node embeddings and role embeddings can be used as features by machine learning algorithms to identify collusion. We create a new network dataset from a large sample of World Bank procurement contracts between 2001 and 2021 for our experimentation. The network dataset has different World Bank entities, borrower countries and suppliers as nodes and awarded contracts as edges. The edges are weighted by contract amount. We also compared our results with the list of suppliers barred from doing business with the World Bank for engaging in various fraudulent activities in order to evaluate the success of our experiment.

1 Introduction

A common way of thinking of financial transactions as networks is to consider entities as nodes and money flows as edges. Fraud in financial transactions is seen as an anomalous activity, that appears as a high-value transaction or a series of unanticipated transactions among several nodes. A simple rule-based anomaly detection on transaction amounts or observing a few nodes for fraudulent transactions can produce a lot of false positives in complex networks, where millions of transactions worth billions of dollars take place. The Bank Secrecy Act requires that any cash transaction of $10,000 or more be reported to the Internal Revenue Service. A simplistic rule like this rarely catches fraudulent activity because the majority of $10,000 and above transactions are valid. To circumvent the system, fraudsters use more sophisticated techniques that are harder to detect. Hence, constant rule discovery to prevent fraud is required.

Procurement fraud is one of the most common types of fraudulent activity. In the procurement process, fraud often occurs when employees involved in purchasing collaborate with suppliers to purchase substandard goods and services, or products that are overpriced in order to receive kickbacks. The public procurement processes are more vulnerable to such collusive fraud due to weak sourcing controls and changing policies with governments or controlling authorities (corrupt influence). An estimated[1] 9.5 trillion US dollars of procurement expenditure occurs globally every year, and corruption is estimated to cost 10 to 25 percent of a public contract's overall value, according to the United Nations Office on Drugs and Crime.

[1] https://www.worldbank.org/en/news/feature/2020/03/23/global-public-procurement-database-share-compare-improve

In procurement processes where internal fraud controls are strong, another fraud pattern referred to as ‘Collusive bidding’ gets more dominant. Antitrust Division of the Department of Justice (Justice, 2021) describe various collusive behaviors in the procurement process:

- **Bid Suppression** – Competitors either do not submit bids or withdraw submitted bids
- **Bid Rotation** – Competitors take turns on contracts
- **Market Division** – Competitors divide territories or customers among themselves

In all such collusive behaviors, collaborators carry out these activities in the strictest of secrecy and usually a whistleblower's allegations accompanied by legal investigations uncover these illegal collusions.

Collusive fraud exists due to direct or indirect relationship between business entities which can also be viewed from the perspective of anomalous communities in a network. This study evaluates OddBall as well as different node embedding methods for detecting collusion in procurement processes.

2 Related Work

Machine learning approaches have been at the forefront of fraud detection in financial transactions related to procurement. (García Rodríguez et al., 2022) in their study examine ability of various ML algorithms to detect collusive fraud in 6 different publicly available procurement datasets. As the legal authorities have investigated these datasets for price fixing and bid rigging in the past, auctions were classified as either ‘collusive’ or ‘competitive’ for modeling. The authors then tested 11 different ML algorithms for their performance on these datasets under various technical settings and found Ensemble methods as the top-performing ML algorithms. Based on error metrics Accuracy, Precision and Recall, they identified Linear models: Stochastic Gradient Descent, Support Vector Machines:, K Neighbors, Multi-Layer Perceptron, Bernoulli and Gaussian Naive Bayes and Gaussian Processes to be worst performing. The best results were obtained using the Extra Trees and Random Forest models, AdaBoost, and Gradient Boosting algorithms. While the utility of machine learning methods in classifying collusive vs competitive bid was proven, the ML models are dependent on label information. This is a major shortcoming of ML methods that is addressed through network analysis approaches.

In their study, (Akoglu et al., 2010) proposed an unsupervised algorithm titled OddBall for detecting anomalous nodes in large networks. Their approach focuses on finding ‘Near-cliques’, ‘Stars’ ‘Heavy vicinities’ characterized by repeating transactions and ‘Dominant heavy links’ which identify strongly connected pairs. Based on 1 step away neighborhood (Egonet), the Oddball approach estimate the parameters for Egonet Density Power Law (EDPL), Egonet Weight Power Law (EWPL) and Egonet λ Power Law (ELWPL). The authors then calculate an outlier score penalizing a node for its deviation from the predicted power law relationships. The outlier score is calculated as “distance to fitting line” and is given by -

$$\frac{max(observed, predicted)}{min(observed, predicted)} * log(|observed - predicted| + 1)$$

This approach limits the minimum outlier score to zero and allows ranking of nodes by their outlier score. While OddBall approach is focused on finding only few anomalous subgraphs structures such as star and clique, (Pei et al., 2020) extend their study further to other subgraph structures that indicate collective fraudulent behavior. Typical examples being ring structures for money laundering or tree structures for pyramid payments. A new subgraph embedding approach is introduced to represent each subgraph as a feature vector, overcoming the challenge of node-level embedding that preserves only the local connections between nodes. The usage of subgraph level embedding instead of node level embedding allows them to preserve the local structure of nodes (intra community edges) as well as the global connectivity patterns (inter community edges) of subgraphs. Following the feature vector conversion, the authors use Isolation Forest as an anomaly detection method to build an ensemble of iTrees and the anomalies are identified as instances with short average path lengths on the iTrees. By detecting pyramid schemes from a transaction network, the authors also demonstrate the effectiveness of their new approach.

3 Data

The World Bank Group (WBG) is a major source of funding for developing countries around the world. These funds may be used for purchasing goods, construction (works), and other non-consulting or consulting services. Several World Bank subsidiaries aid in this procurement process, including the International Bank for Reconstruction and Development (IBRD) and International Finance Corporation (IFC). The World Bank publishes regular data on the contracts awarded to various vendors across the world through its open data platform[2].

For the purpose of our analysis, we use the major contract award dataset[3] published by the World Bank. This dataset provides us the details of burrower country, supported project details, its objective, supplier to whom contract is awarded and the total contact award amount in USD. The dataset includes information about 162K contracts issued between 2001 and 2022.

The World bank also publishes the list[4] of suppliers it debars from the procurement process due to indulgence in fraud and other corrupt activities. We utilize this dataset to validate the findings of our analysis. The key attributes of the data published by the World bank are shown in the table below:

#	Attribute	Description
1	Fiscal Year	The fiscal year in which contract was raised
2	Borrower Country	The country which took the loan from World Bank
3	Supplier	The supplier firm which was given the contract to complete the project in borrower country
4	Supplier Country	The country in where supplier is legally registered
5	Total Contract Amount (USD)	The total contract amount in US Dollars
6	Borrower Contract Reference Number	Unique identifier for the contract

Table 1: Key Attributes in Procurement Dataset

For our analysis we considered both bipartite and unipartite graphs. In bipartite graph borrower country and supplier are the two groups of nodes with contract amount as edges while in unipartite chart we only use supplier as nodes. The unipartite chart is created as a projection of bipartite chart and below are the steps we used to create the unipartite projection network:

Step1: Using text analytic approaches we clean the supplier and borrower text fields in our dataset and remove text impurities like punctuations, links, special characters etc. This step is necessary to do an effective fuzzy match with the suspended supplier list and avoid duplicate node with minor variation in name

Step2: We create a bipartite graph using the supplier , borrower country and the contract amount as weights for each fiscal year

Step3: Lastly, we create a weighed unipartite projection of borrower country onto supplier country such that the edges equal the sum of contract value

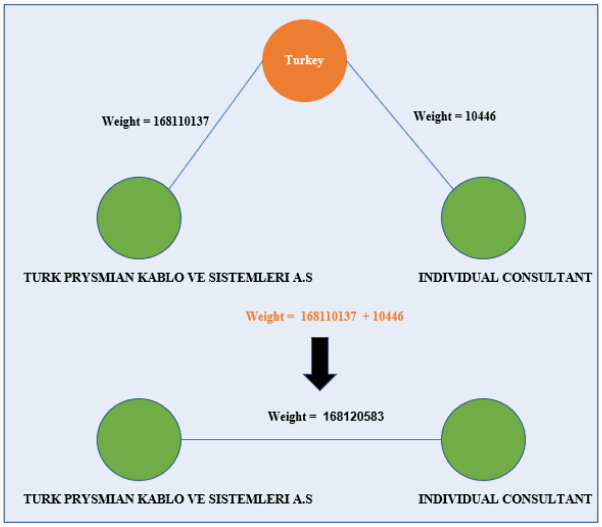
Turkey
Weight = 168110137
Weight = 10446
TURK PRYSMIAN KABLO VE SISTEMLERI A.S
INDIVIDUAL CONSULTANT
Weight = 168110137 + 10446
Weight = 168120583
TURK PRYSMIAN KABLO VE SISTEMLERI A.S
INDIVIDUAL CONSULTANT

Figure 1: Conversion of Bipartite Network to Unipartite Network

[2] World Bank Open Data
[3] Major Contract Awards
[4] Procurement - World Bank Listing of Ineligible Firms and Individuals

Finally, we have a network of each fiscal year where suppliers are the nodes and edges are weighted on total contract amount. We follow the same steps for creating multi-year unipartite networks.

4 Methods

4.1 OddBall

The algorithm OddBall identifies several power law patterns governing the ego-nets of all nodes, i.e., subgraphs of all nodes and their neighbors, and uses them for anomaly detection. OddBall computes a Least Squares fitting line for a power law first, and then measures the anomalousness of each node by its distance from the fitting line. It uses the number of neighbors of ego (N), number of edges in egonet (E), total weight of egonet (W) and principal eigenvalue of the weighted adjacency matrix of egonet (λw) as features. The anomalies are then identified on different pairs of regression line of these features:

- Clique Star (E vs N) : near-cliques and stars
- Heavy Vicinity (W vs E) : recurrences of interactions
- Dominant Pair (λw vs W) : single dominating heavy edge (strongly connected pair)

The underlying assumption around the distributions of features is listed below:

1. number of nodes N_i and the number of edges E_i follow a power law

$$E_i \propto N_i^{\alpha}, 1 \leq \alpha \leq 2.$$

2. total weight W and the number of edges E follow a power law

$$W_i \propto E_i^{\beta}, \beta \geq 1$$

3. the principal eigenvalue λw of the weighted adjacency matrix and the total weight W follow a power law

$$\lambda_{w,i} \propto W_i^{\gamma}, .5 \leq \gamma \leq 1$$

4. the rank R_{ij} and the weight W_{ij} of edge j follow a power law

$$W_{i,j} \propto R_{i,j}^{\theta}, \theta \leq 0$$

4.2 Node2Vec Embedding

Node2vec embedding, a representation technique formulated by (Grover & Leskovec, 2016) is a method for encoding structural information about a graph by teaching it a mapping that embeds the nodes from a graph as points in a low-dimensional vector space $\mathbb{R}^d$. This mapping yields geometric relationships that reflect the original graph's structure in this learned vector space. The Node2Vec approach uses random walks statistics to learn the node embeddings instead of using deterministic distance measure. The decoder function to identify the embedding is given by

$$DEC(z_i, z_j) \triangleq \frac{e^{z_i^{\gamma} z_j}}{\sum_{v_k \varepsilon V} e^{z_i^T z_k}} \approx p_{G,T}(v_j | v_i)$$

Where $p_{G,T}(v_j | v_i)$is the probability of visiting v_j on a length-T random walk starting at v_i, with T defined to be in the range T ∈ {2,..., 10}. The node2vec algorithm introduces two random walk hyper-parameters, p, and q, which bias the random walk. The hyper-parameter p determines whether the walk will immediately revisit a node, while q dictates whether the walk will revisit a node's immediate neighborhood. Tuning these parameters allows us to choose between learning community structures or embeddings that emphasize local structural roles.

As part of our study, we were able to use the learned embeddings as features to perform two machine learning tasks - find the most similar set of suppliers sanctioned by the World Bank, and use the embeddings as features for Isolation Forest, a machine learning algorithm for detecting outliers.

4.3 Role Embedding

Node embedding entails encoding nodes as low-dimensional vectors that summarize graph positions and the structure of their local graph neighborhood, while role embedding allows us to learn representations that encapsulate nodes' structural roles, independently of their position within the global graph. Similar to our approach with Node embeddings, we used Role embeddings as features for Isolation Forest algorithm to detect

outliers. Though Struc2Vec (Ribeiro et al., 2017) and GrpahWave (Donnat et al., 2017) are the more popularly used algorithms for role embeddings, we decided to experiment with a relatively new approach RiWalk conceived by (Ma et al., 2019).
RiWalk separates the structural embedding problem into two steps: role identification and network embedding. Role identification is based on the idea of substructure and relabeling developed for graph kernels that compute similarity between subgraphs based on similarity between their components. Following this intuition, the methodology assumes that if two nodes exhibit similar distributions of degrees of their neighbor nodes, they are structurally similar. It further uses distances between context nodes and anchor nodes in its role identification function. After role identification a skip-gram model with negative sampling is used to learn network embedding.

As an alternative to using node embedding and role embedding separately, in our last experimentation we used both node embedding and role embedding as features for our machine learning algorithm to detect outliers.

We used the code available on the GitHub repositories oddball_py3[5] and RiWalk[6] for implementation in our project. In addition, we also utilized python packages ‘Node2Vec’ and ‘graphrole’ for generating node embedding and role assignment.

5 Results

5.1 OddBall Outliers

We executed the OddBall model on unipartite network structure for year 2021 and below table lists down the top 5 suppliers in each anomaly category along with the outlier score.

Supplier	**Score**
Tractebel Engineering S.A.	9.71
Grant Thornton	9.62
Ambitious Construction Company Ltd	9.31
Techno Three (U) Ltd	9.31
Sinohydro	9.11

Table 2: Year 2021 - Anomaly Type : Star

Tractebel Engineering S.A.[7] , a star structure identified by the OddBall algorithm with the highest outlier score, has 92 other suppliers as its first neighbors. A French listed entity, it had major contracts in Africa and Asia. According to the World Bank suspension list, this entity has been blocked from future business under the category of "Cross Debarment: IDB". In this category, firms and individuals are sanctioned for engaging in fraudulent, corrupt, collusive, coercive, or obstructive practices by the World bank.

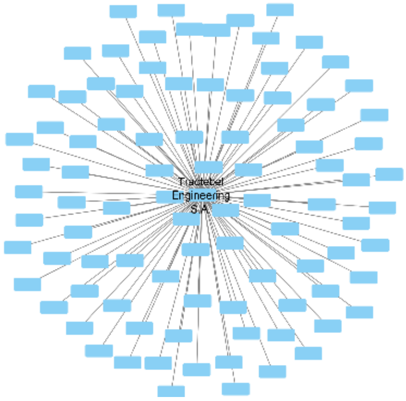

Figure 2: Tractebel Engineering S.A., a Star Anomaly

We also looked into the history of Tractebel Engineering SA and found that this firm has contracts every alternate year since 2013. This behavior looks suspicious and can be accounted under “bid rotation” pattern that we discussed in section 1.

Supplier	**Score**
Trademaster Resources	70.73
Isecure Integrated Business Solutions	70.73
Moderna Switzerland Gmbh	70.73
Tractebel Inc. Resources Environment And Economics Center For Studies Inc.	70.73
Macro Machinery Industrial Supply Corp./Ferdstar Builders Contractors (Jv)	70.73

Table 3: Year 2021 - Anomaly Type : Dominant Edge

Tractebel Inc., a related entity to Tractebel Engineering also appears in our top 5 outliers

[5] OddBall Python Implementation
[6] RiWalk Python Implementation
[7] Tractebel: Sanctioned List

results for anomaly type dominant edge. From figure 4 we can notice a high weighted edge between Moderna Switzerland and Tractebel Inc.

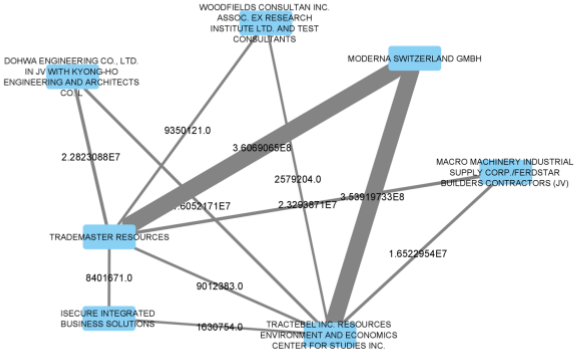
WOODFIELDS CONSULTAN INC. ASSOC. EX RESEARCH INSTITUTE LTD. AND TEST CONSULTANTS
MODERNA SWITZERLAND GMBH
DOHWA ENGINEERING CO., LTD. IN JV WITH KYONG-HO ENGINEERING AND ARCHITECTS CO.L
9350121.0
3.6069065E8
2.2823088E7
2579204.0
MACRO MACHINERY INDUSTRIAL SUPPLY CORP./FERDSTAR BUILDERS CONTRACTORS (JV)
3.53919733E8
.6052171E7
2.3293871E7
TRADEMASTER RESOURCES
1.6522954E7
8401671.0
9012383.0
ISECURE INTEGRATED BUSINESS SOLUTIONS
1630754.0
TRACTEBEL INC. RESOURCES ENVIRONMENT AND ECONOMICS CENTER FOR STUDIES INC.

Figure 3: Dominant Edge Identified For Supplier Trademaster Resources (Weighted by Contract Amount)

In fraud investigations, heavy vicinity neighborhood can provide a smaller sample size to begin investigation as these entities have large interactions among themselves, and fraudsters often prefer to deal with multiple subsidiaries or companies incorporated abroad. Taking a closer look at the transactions, origin of companies doing business (same registration address) or corporate family tree can reveal hidden collusion.

Supplier	Score
Proshika Manobik Unnayk Kendra	97.14
Nature Conservation Management (Nacom)	97.14
Hg Power Transmission Sdn. Bhd.	97.14
Joint Venture Of Dev Consultants Limited	97.14
Kranti Associates Ltd.	97.14

Table 4: Year 2021 - Anomaly Type : Heavy Vicinity

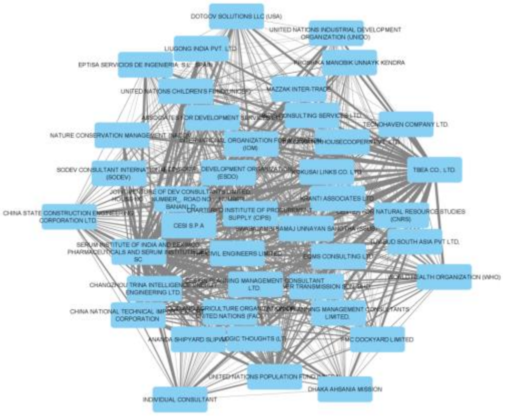
DOTGOV SOLUTIONS LLC (USA)
UNITED NATIONS INDUSTRIAL DEVELOPMENT ORGANIZATION (UNIDO)
LIUGONG INDIA PVT. LTD
EPTISA SERVICIOS DE INGENIERIA S.L. SPAIN
PROSHIKA MANOBIK UNNAYK KENDRA
UNITED NATIONS CHILDREN'S FUND (UNICEF)
MAZZAK INTER-TRADE
TECNOHAVEN COMPANY LTD.
NATURE CONSERVATION MANAGEMENT (NACOM)
SODEV CONSULTANT INTERNATIONAL (SODEV)
TBEA CO., LTD
KOKUSAI LINKS CO. LTD
KRANTI ASSOCIATES LTD
CHINA STATE CONSTRUCTION ENGINEERING CORPORATION LTD
CHARTERED INSTITUTE OF PROCUREMENT SUPPLY (CIPS)
CESI S.P.A
SERUM INSTITUTE OF INDIA AND BEXIMCO PHARMACEUTICALS AND SERUM INSTITUTE OF INDIA SC
EGMS CONSULTING LTD
WORLD HEALTH ORGANIZATION (WHO)
CHANGZHOU TRINA INTELLIGENT ENERGY ENGINEERING LTD
CHINA NATIONAL TECHNICAL IMPORT CORPORATION
ANANDA SHIPYARD SLIPWAY
LOGIC THOUGHTS (LT)
FMC DOCKYARD LIMITED
UNITED NATIONS POPULATION FUND
DHAKA AHSANIA MISSION
INDIVIDUAL CONSULTANT

Figure 4: Heavy Vicinity Neighborhood

5.2 Node2Vec Embedding (Most Similar Nodes)

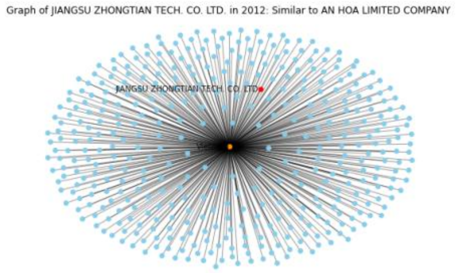
Graph of JIANGSU ZHONGTIAN TECH. CO. LTD. in 2012: Similar to AN HOA LIMITED COMPANY
JIANGSU ZHONGTIAN TECH. CO. LTD.

Figure 5: Most Similar Supplier (AN HOA Limited)

In node embedding, nodes are encoded in such a way that the similarity in the embedding space (e.g., dot product) approximates the similarity in the original network. The same characteristic was used to identify firms similar to those debarred by the world bank. One such example is AN HOA Limited, which was debarred by the World Bank in 2016. The most similar function returned Jiangsu Zhongtian Technology Co., Limited[8] which was sanctioned in 2019 under the same procurement guidelines, 1.14(a)(ii). The result indicates that a large number of supplier firms engage in similar malpractices. We can be proactive by investigating the firms that appear on most similar lists of debarred firms. Fraud is always hidden, and it takes creative analytics to uncover it.

5.3 Role Embedding (Isolation Forest)

We generated role embedding for bipartite network for the time period 2010 -2021. The experiment we performed using just role embedding as features for outlier detector algorithm Isolation Forest enabled us to identify a large number of outlier suppliers that mainly act as bridges and hubs, i.e., supply products or provide services to multiple countries.

[8] Jiangsu Zhongtian

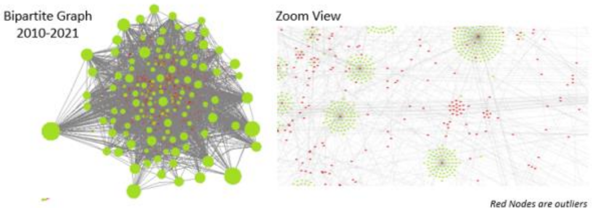

Figure 6: Outlier - Bipartite Network (2010 -2021)

Even though the results were not very encouraging, we used the learning from these results to combine the role embeddings and node embeddings features to develop a new outlier detection model.

5.4 Role and Node Embedding (Isolation Forest)

Based on the combination of node and role embedding as input features, Isolation Forest identified an anomalous cluster of 224 firms in the Bipartite graph which spans the years 2016 – 2021.

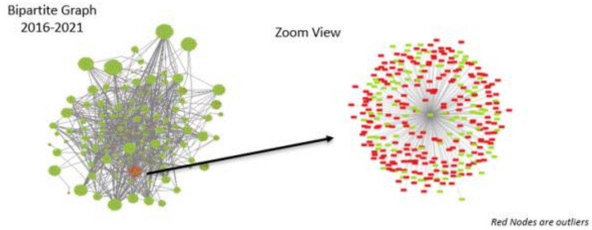

Figure 7: Outlier - Bipartite Network (2016 -2021)

This cluster was associated with Western Africa, and in our limited investigation of 224 anomalous suppliers, we found one supplier SINOHYDRO[9], which was sanctioned by the World Bank.

6 Discussion & Conclusion

As part of our exploration of multiple approaches to identify anomalies in networks, we were able to demonstrate that outliers identified by our approaches were involved in fraudulent activities as they appeared in the World Bank's sanction list. We found that network analysis techniques can capture the behavior and interaction of entities generated by financial transactions. We did not add any additional features to the data, instead we extracted network topology and role information from the nodes based on various embedding techniques. Overall, the results are promising and can be further explored by tuning the hyperparameters as well as taking into account the time periods we excluded from our analysis due to inability to work with a larger network. With millions of edges in the unipartite network we built over the period 2010-2021, our algorithms failed to process it efficiently. Therefore, we would like to explore how multi-processing can be used to work on larger graphs and produce more holistic results.

Moreover, we would also like to delve into different node embedding techniques that are more suitable for bipartite networks, as well as experiment with normalized edge weights, which is not part of our current approach.

References

Akoglu, L., McGlohon, M., & Faloutsos, C. (2010). *OddBall: Spotting anomalies in weighted graphs* (Vol. 6119 LNAI) [Conference Paper]. https://doi.org/10.1007/978-3-642-13672-6_40

Donnat, C., Zitnik, M., Hallac, D., & Leskovec, J. (2017). Spectral Graph Wavelets for Structural Role Similarity in Networks. *ArXiv*, *abs/1710.10321*.

García Rodríguez, M. J., Rodríguez-Montequín, V., Ballesteros-Pérez, P., Love, P. E. D., & Signor, R. (2022). Collusion detection in public procurement auctions with machine learning algorithms [Review Article]. *Automation in Construction*, *133*. https://doi.org/10.1016/j.autcon.2021.104047

Grover, A., & Leskovec, J. (2016, 2016 / 08 / 13 /). Node2vec: Scalable feature learning for networks.

Price fixing, bid rigging, and market allocation schemes: what they are and what to look for, (2021). https://www.justice.gov/atr/file/810261/download

Ma, X., Qin, G., Qiu, Z., Zheng, M., & Wang, Z. (2019). RiWalk: Fast Structural Node Embedding via Role Identification. In (pp. 478-487): IEEE.

Pei, Y., Lyu, F., Ipenburg, W., & Pechenizkiy, M. (2020). *Subgraph anomaly detection in financial transaction networks*. https://doi.org/10.1145/3383455.3422548

Ribeiro, L. F. R., Saverese, P. H. P., & Figueiredo, D. R. (2017, 2017 / 08 / 13 /). Struc2vec: Learning node representations from structural identity.

[9] Sinohydro