Improving Public Procurement Collusion Detection With Graph-based Machine Learning Methodologies

Everton Schneider dos Santos[1], **Márcio Castro**[1], **Jonata Tyska Carvalho**[1]

[1] Postgraduate Program in Computer Science
Federal University of Santa Catarina (UFSC)
P.O. Box 476 – 88.040-370 – Florianópolis – SC – Brazil

e.schneider.s@posgrad.ufsc.br,

marcio.castro@ufsc.br, jonata.tyska@ufsc.br

Abstract. *Public procurement is a complex process often susceptible to corruption and Machine Learning (ML) has emerged as a promising approach to identifying fraud in public procurement. While many ML methods for fraud detection rely on tabular data, information from the network of relationships between the entities involved in procurement process remains underutilized. This work aims to fill this gap with a study of ML methodologies for detecting collusion in public procurement using data extracted from the relationships network. Enriching the "Operation Car Wash" with topological information extracted from graphs helped improve the results by 1% and decrease the variability of the evaluated models by almost 5%.*

1. Introduction

Public procurement is a structured competitive bidding process that governments widely employ to allocate public funds for acquiring essential goods and services [Curtis and et al. 1973]. This mechanism is intended to promote transparency, efficiency, and fairness, ultimately ensuring that governments and taxpayers obtain greater value for money [García Rodríguez and et al. 2022]. Despite these goals, public procurement is inherently complex and is often perceived as a significant source of corruption [Sanz et al. 2024].

The identification of corruption in public procurement has received widespread academic attention due to its central role in economic growth [Fazekas and Wachs 2020]. Machine Learning (ML) models have been widely used in the literature. *Gallego et al.* used contract data to train ML models capable of predicting if a public procurement process will result in a corruption investigation or breach of contract [Gallego et al. 2021]. *Wallimann and Sticher* used a public procurement dataset from the Swiss railway infrastructure market to build ML models capable of detecting collusion [Wallimann and Sticher 2023].

Although the network of relationships between entities is central to the public procurement process, it remains an underutilized data source in fraud detection research, as shown in [Schneider dos Santos et al. 2025]. This work aims to address this gap by conducting a study on the use of topological information extracted from a graph of bids and tenders, aiming to enrich a dataset of collusive tenders and evaluate whether this approach can improve the collusion detection rate of machine learning models. In this

study, we also trained Deep Learning (DL) models capable of processing graph data and evaluate their ability to detect collusion.

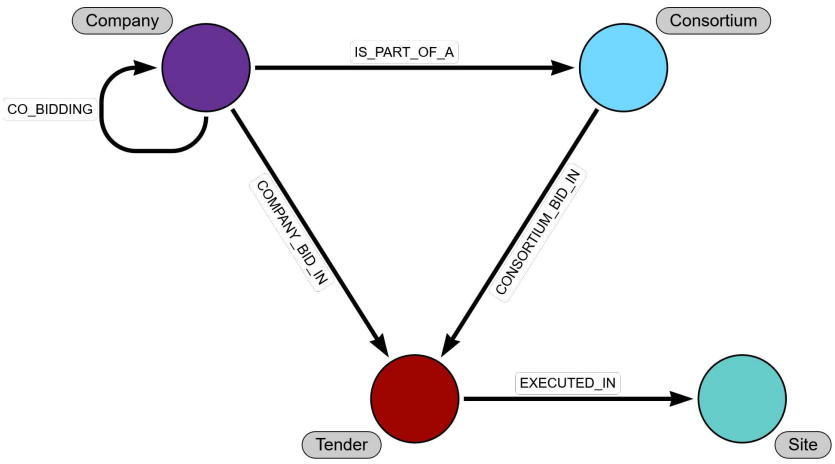
Using the "Operation Car Wash" dataset enriched with topological information extracted from the network of relationships between the entities involved in the procurement process, the results from traditional ML methods improved by 1% and the variability in performance using different data split techniques decreased by nearly 5%.

2. Data and Experimental Setup

The data used in this work contains information about collusive tenders in the context of the "Operation Car Wash". The dataset also contains screening variables, statistical features about bids in a tender that can be used to improve the collusion detection rate of ML methods. This dataset was made available by [García Rodríguez and et al. 2022] and enriched by [dos Santos et al. 2024] with information about the companies participating in the tenders, including: legal nature, capital invested, and the founding date.

Following [dos Santos et al. 2024], our first methodology involved training traditional ML models using the "Operation Car Wash" dataset in a tabular format. The scikit-Learn [Pedregosa and et al. 2011] library was used to implement the following models: GradientBoosting (GB), RandomForest (RF) and the Multi-layer Perceptron (MLP). In the second methodology, we created a heterogeneous graph based on the information from the "Operation Car Wash". The proposed model is shown in Figure 1.

Figure 1. Graph modeling of the "Operation Car Wash" dataset



The proposed graph is designed to train ML models for graph classification problems. In these applications, models are trained using a dataset of independent graphs to make predictions specific to each graph [Hamilton 2020]. We used this model to extract relevant topological information from the bidding network. Features such as centrality measures and node embeddings were calculated and used to enrich the tabular dataset. The objective of this study is to evaluate the efficacy of this enrichment in enhancing the performance of models trained with tabular data.

The results of traditional and graph models were assessed using the balanced accuracy metric. We also executed the following data split strategies: Repeated Holdout with 100 repetitions, Repeated K-Fold with 10 repetitions, and Nested K-Fold. A hyperparameter tuning process was conducted to find an optimal architecture for the models.

3. Results and Discussion

The first strategy for enriching the tabular dataset consists of calculating centrality measures for all nodes in the bidding network. The following centrality measures were calculated: Betweenness, Eigenvector, and PageRank. The second enrichment strategy consists of generating node embeddings using the Node2vec algorithm. These embeddings were based on past co-bidding relationships, which were modeled as the `CO_BIDDING` relationship shown in Figure 1. We used the Neo4j Graph Data Science library[1] to create these embeddings and centrality measures. Table 1 presents the models that achieved the best results with the enriched datasets.

Table 1. Results of the enrichment process using graph-based features

	Centrality Measures		Node Embeddings	
Data Split	**Model**	**Balanced Acc.**	**Model**	**Balanced Acc.**
Holdout	GB	88,16 (8.22)	GB	86.58 (8.92)
K Fold	RF	**90,53 (6.81)**	GB	**87.86 (9.51)**
Nested CV	RF	87,94 (11.10)	GB	87.12 (12.86)

As shown in Table 1, enriching tabular data with graph information enhanced model performance over the baseline strategy. The addition of centrality measures increased model accuracy by 1% and decreased the standard deviation of the results by nearly 5%. This suggests the proposed enrichment strategy reduces performance variability, thus enabling a more reliable selection of models for detecting public procurement collusion.

We also used the proposed graph modeling, shown in Figure 1, to train a Graph Convolutional Network (GCN) for a graph classification task. In this task, a model is trained with the objective of predicting whether a tender is collusive or non-collusive. The model was built using stacked GraphSage layers in the PyTorch Geometric (PyG) library[2]. Table 2 shows the results.

Table 2. Results of the GCN model for graph classification

Strategy	**Balanced Accuracy**	**Std**
Baseline	71.15	12.75
K-Fold	75.24	7.58
Nested CV	**77.06**	12.03

Despite the use of a hyperparameter optimization process, the implemented strategies did not significantly improve the model's accuracy. A potential explanation for these results is the limited size of the dataset used. The "Operation Car Wash" dataset is composed of 100 tenders, of which only 32 are classified as collusive. The limited number of samples complicates data splitting, particularly for cross-validation.

[1] https://neo4j.com/docs/graph-data-science/current/algorithms/

[2] https://pytorch-geometric.readthedocs.io/en/latest/

4. Conclusion

This work investigated the use of graph-based data to improve the collusion detection rate of different ML methodologies. The enrichment of tabular data with topological information, extracted from the network of relationships between the entities involved in the procurement process, not only improved the accuracy of traditional models but also significantly reduced the variability of results from different data splitting techniques.

The use of graph-based DL models, however, did not outperform conventional methods that employed tabular data from the collusive tenders of the "Operation Car Wash". Further experiments are needed to validate the efficacy of GCNs for detecting collusion in public procurement. Future work should include using larger and richer datasets and testing alternative convolutional layers. Future work could also explore node classification to assess its viability for this problem, especially with limited data.

Acknowledgments. This research was funded by the Coordination for the Improvement of Higher Education Personnel Foundation (CAPES) and the Public Prosecutor's Office of Santa Catarina (MPSC).

References

Curtis, F. and et al. (1973). Closed competitive bidding. *Omega*, 1(5):613–619.

dos Santos, E. S., dos Santos, M. M., Castro, M., and Carvalho, J. T. (2024). Performance variability of machine learning models using limited data for collusion detection: A case study of the brazilian car wash operation. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 431–443, Porto Alegre, RS, Brasil. SBC.

Fazekas, M. and Wachs, J. (2020). Corruption and the network structure of public contracting markets across government change. *Politics and Governance*, 8(2):153–166.

Gallego, J., Rivero, G., and Martínez, J. (2021). Preventing rather than punishing: An early warning model of malfeasance in public procurement. *International Journal of Forecasting*, 37(1):360–377.

García Rodríguez, M. J. and et al. (2022). Collusion detection in public procurement auctions with machine learning algorithms. *Automation in Construction*, 133:104047.

Hamilton, W. L. (2020). *Graph representation learning*. Morgan & Claypool Publishers.

Pedregosa, F. and et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Sanz, I. P., Iturriaga, F. J. L., and Blanco-Alcántara, D. (2024). A neural network approach for predicting corruption in public procurement. *European Journal of International Management*, 22(2):175–197.

Schneider dos Santos, E., Machado dos Santos, M., Castro, M., and Tyska Carvalho, J. (2025). Detection of fraud in public procurement using data-driven methods: a systematic mapping study. *EPJ Data Science*, 14(1).

Wallimann, H. and Sticher, S. (2023). On suspicious tracks: machine-learning based approaches to detect cartels in railway-infrastructure procurement. *Transport Policy*, 143:121–131.