Graph Data Mining for Detecting Collusions in Bidding Processes: A Case Study

André Ormastroni Victor[1], Luís A. M. Sales[1], Rodrigo S. Moreira[1], Carlos E. C. de Moraes[1], Luis G. Lima[1], João F. Rocha[1], Bruno S. N. Contursi[1], Thiago Meirelles[1]

[1]Investigative Intelligence (IEMSI/II) – Petróleo Brasileiro S.A., Rio de Janeiro - RJ - Brazil.

{aovictor, luis.sales, rodrigo.moreira, carloseduardo.moraes, lg_bessalima, joao.felipe, bruno.contursi, thiago.meirelles}@petrobras.com.br

Abstract. *Bidding collusion causes significant financial damage in Brazil and compromises the efficiency of public service delivery. This work proposes the representation of bidding disputes as a directed graph and the use of data mining techniques on graphs to detect suspicious data patterns that may reveal anti-competitive behaviour. The algorithms and techniques used were tested in a case study with the real bidding data of a Brazilian mixed economy company. The results show that to ensure the integrity of the bidding processes it is required to develop more complex preventive controls which check more comprehensively the entire network of relationships in the graph.*

1. Introduction

Bidding is the main administrative process to which companies wishing to provide goods and services for any public administration body (direct or indirect) in Brazil are subjected to, competing under the same rules so that the proposals that best meet public interest are selected (Costa, 2019). According to data from the Brazilian public procurement portal, of the little more than R$ 258 billion transacted in 2023, only 0.15% occurred with exemption or non-requirement of this instrument (Brazil, 2024).

Therefore, ensuring the integrity and competitiveness of bidding processes is essential for fairness, efficiency, and transparency in the use of public resources. However, the Association of Certified Fraud Examiners in its "Report to the Nations" (ACFE, 2024) reveals an alarming reality: it is estimated that 5% of organisations' annual revenue was lost to fraud and corruption schemes in 2023. In the energy sector, for instance, the median losses reached the figure of $152,000 per case.

In this context, this paper aims to present the application of detection methods for possible collusive schemes among bidding participants and continue the work initiated on Victor *et al.* (2024). The adopted approach involves modelling data in a property graph database and using graph data mining techniques and algorithms to identify outliers. The case study consists of the bidding database from 2006 to 2022 of Petrobras, a mixed economy company in Brazil. The originality and contribution of this study lie in the application of graph-based algorithms to a real case, offering a new approach to the prevention and combating of fraud in bidding processes.

This article is organised into five sections, including this introduction. Section 2 presents as a brief literature review on how the topic is approached in the bibliography.

Section 3 presents the methodological approach and the organisation of data in the format of a property graph database. Section 4 presents the analysis and interpretation of the proposed indicators. Finally, Section 5 summarises the main contributions of the article, pointing out limitations and future research paths.

2. Related works

The present paper expands the previous work published on Victor *et al.* (2024) and shares with it the bibliography corresponding to the intersection between data mining techniques based on graphs with its application on the detection of fraud schemes. Paulo and Rodrigues (2022) conducted an in-depth analysis of TCU data, resulting in the development of a graph-based method for visualising and identifying irregularities in bids. Padhi and Mohapatra (2011) distinguished themselves by utilising statistical techniques to detect collusion in India's governmental procurement processes. Similarly, Liu *et al*. (2016) applied graph modelling but concentrated on uncovering suspicious connections in the U.S. healthcare industry, especially between physicians and pharmaceutical firms. Silva *et al*. (2020) used data mining techniques to detect suspicious associations among bidding companies in the Armed Forces. Another noteworthy work is by Amaral (2020), who developed community detection techniques in graphs to identify communities of companies with suspicious activities in public procurement in the State of Rio de Janeiro.

Other than the related works, this work builds on the previous one as it harnessed not only the dataset, but also the objective of finding possible collusion schemes among bidders. In the original work, indicators of fraud were searched via the commonality of paths in the graph that were unexpected, considering a perfectly competitive environment among bidders. For instance, one would be suspicious if the same user modified bidding proposals of different competitors (cf. the node Metadata on Figure 1). The main improvement and originality of the present paper is that it applies a richer and more graph-oriented set of algorithms, which will be detailed in the following sections.

3. Methodology

3.1. Schema

As per the Figure 1, the property graph schema was constructed considering the network of relationships formed during the lifecycle of the bidding process in the source systems. The identification of a procurement need leads to the creation of a bidding opportunity (node Opportunity) published on the company's procurement portal, detailing the items (node Item) to be bid on (relationship CONTAINS). Interested suppliers (node Company) register on the site, linking a responsible user (node User) for operations on the portal (relationship REGISTERED).

After registration, suppliers can participate in the bidding process by placing bids for the opportunity items (relationship BIDS). Once the bidding period ends, a final version is consolidated into a unit price spreadsheet (node UPS_File), which has (relationship MODIFIED_BY) as metadata the user who performed the last modification (node Metadata) and is submitted and presented according to the opportunity and the proposing company (relationship PRESENTS). The analysis of the bids determines the winner of the bidding process, establishing the WINS relationship, and also the Pearson’s correlation among the bidden unit prices between two suppliers in an opportunity

(relationship COLIN). Additional relationships such as COMPETES_FOR, JOINS, and CREATES are derived from other existing relationships and were created to facilitate queries during the analysis.

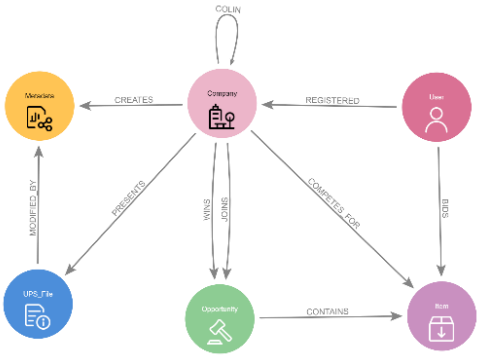

Figure 1. Graph schema.

3.2. Algorithms and Indicators of Collusion

From the graph instantiated after the extraction, transformation, and loading process, indicators were developed to detect suspicious cases of collusion among the supplier nodes that form in the subgraphs of the bidding disputes. In this regard, we assume the premise that collusion occurs through the simulation of non-existent competition (OECD, 2021).

This simulation can occur in two distinct scenarios. The first scenario involves a shell company, which exists only on paper and does not operate in the market. Its purpose is to simulate non-existent competition, allowing the other bidder to win the contract. The second scenario involves active market companies colluding with pre-arranged prices so that one of them wins the contract. In this case, it may be observed an unlikely statistical coincidence that the companies always compete together in the same set of bids, or there is a high similarity of connections in the network nodes. Both scenarios lead to the development of indicators calculated according to the following algorithms:

- Indicator AR (Association Rule): Widely used in product recommendation applications, the association rules method seeks to reveal significant antecedent and consequent relationships between sets of items. However, instead of suggesting the most relevant goods based on the potential buyer's shopping basket, as in the work of Silva *et al*. (2020), it aims to find unusual associations between companies bidding on the same opportunity items to identify suppliers that simulate competition.
- Indicator CD (Community Detection): The objective of analysing this indicator is to identify groups of suppliers with suspicious behaviour based on the affinity of the relationships they exhibit among themselves. The Louvain algorithm was applied in different configurations to detect communities in networks that group companies based on their relationships, such as participation in opportunities (JOINS), bidding on items (COMPETES_FOR), generation of common modifier user metadata (CREATES) and linear price covariation (COLIN).
- Indicator JS (Jaccard Similarity): The Jaccard similarity index was used between the set of items from suppliers competing for items in a bid, with the size of their

baskets normalised. Being a measure that aims to assess the concordance between two sets, its goal is to identify abnormal cases of item basket similarity between bidders.

It is important to note that all indicators evaluate the behaviour of suppliers based on the entire network of connections and bids in the graph, not just a specific dispute. Therefore, outlier cases represent highly suspicious conditions because they indeed identify recurring behaviours or situations, rather than isolated incidents as some traditional prevention controls are implemented (Rodrigues, 2022).

4. Results

The core of the data being analysed comes from private corporate databases, whose sensitivity and corporate or legal restrictions prevent the full or partial disclosure of the results achieved. In this regard, the present results analysis section will focus not on presenting the total values found by the indicators described in Section 3.2, but rather on relative values within a larger universe, either through percentage representation or distributions, according to the methodological requirements of the indicator in question.

4.1. Indicator AR

In Figure 2, the distributions of the support, confidence, and lift metrics for the association rules of companies in the same opportunities are plotted. The distribution of support values reveals that most rules have low support, which is expected given the large volume of combinations between suppliers. The distribution of confidence values is less concentrated and more uniform, although it still skews towards values below the average. These rules indicate a low association between the suppliers in question, but higher confidence values do not necessarily imply a positively expected relationship between antecedent and consequent baskets

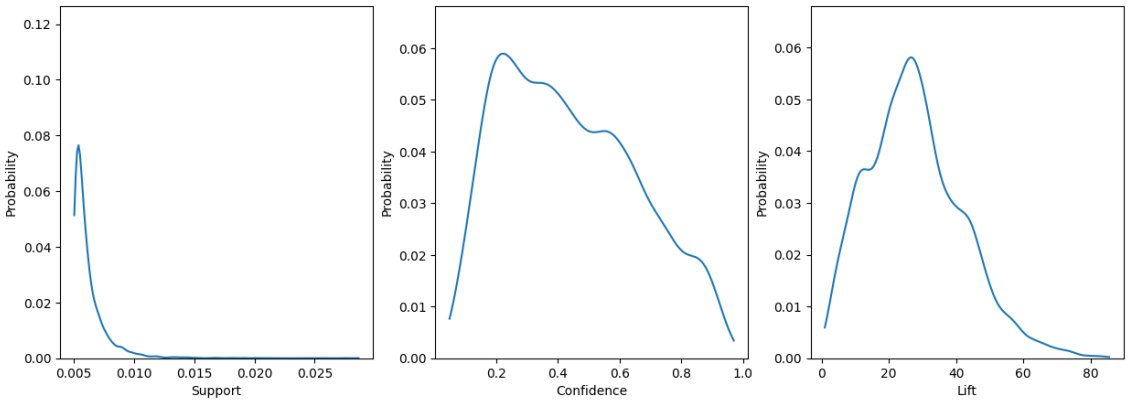

Figure 2. Distribution of support, confidence and lift of the association rules.

Considering the limitations of the confidence metric, the lift metric was used to identify cases of interest. Although only 0.025% of the rules had a lift value less than 1, indicating that almost all rules are relevant to the study, the distribution showed such asymmetry and concentration that extreme values emerged. Suspicious cases were those rules whose lift value violated the interquartile range rule of the distribution, i.e., rules with a lift value greater than 150% of the difference between the third and first quartiles of the distribution.

It should be noted that in the implementation used, the Apriori algorithm explores the space of possible combinations between the suppliers to calculate evaluation metrics. Due to the enormous volume of data, it was necessary to establish a minimum support threshold of 0.5% to perform the calculations, which still represents a significant number of cases of supplier co-occurrence.

4.2. Indicator CD

The execution of the Louvain algorithm generated thousands of communities for each relationship configuration. Therefore, in this indicator, the focus of the analysis was not to individually evaluate the suppliers of each generated community, but rather to identify a group of companies that were classified into the same communities regardless of the relationship used by the algorithm, which indicates a similarly suspicious behaviour among them.

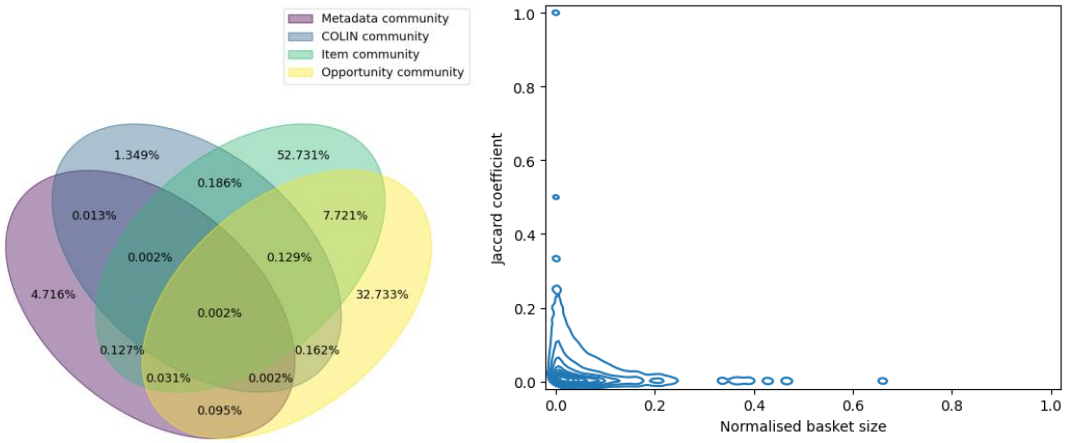

Figure 3. Venn Diagram of the intersections of all community formation criteria indicating the overall proportion of pairs belonging to the communities (left) and Histogram of the distribution of Jaccard similarity and the normalised item basket size of suppliers (right)

Figure 3 (left) shows the Venn diagram with the percentages of pairs of companies for each community configuration and their intersections, compared to all pairs of companies formed by those with some bidding dispute. Two configurations of exclusive community formation criteria were the ones that most concentrated pairs of companies in the same communities out of the total universe of cases: communities formed by companies bidding for the same item (60.93%) and communities of companies bidding for the same opportunities (40.875%).

These groups, when analysed in isolation, do not represent a high risk of suspicion. For example, consider two competing companies in the same commercial sector. It is natural for these companies to be connected with common edges in bids and items, as both are competing in the market. On the other hand, when these companies are also grouped together in the same communities established by other more suspicious criteria, it makes their behaviour more atypical and unlikely. Therefore, the groups that belong to the intersections in the Venn diagram represent the cases with the highest potential risk and should be prioritised for a more thorough investigation, while they also represent a very small percentage of the total universe of companies. Note, for instance,

that only 0.002% of pairs of company vertices were grouped in the same community regardless of the relationship adopted in the execution of the algorithm.

4.3. Indicator JS

While indicator AR targets cases where companies have a high correlation of joint activity in the same bids, indicator JS aims to detect suppliers that exhibit very similar behaviour in the basket of items they bid for. It is natural for companies in the same competitive field to bid together for items in the same tender. However, even in cases of free competition between reputable and independent companies, it is natural for competition to vary across different bidding opportunities. On the other hand, companies that seek to simulate non-existent competition may exhibit high similarity in the items they bid for, as they are always bidding together for the same basket of items. The graph in Figure 3 (right) presents the histogram of the distribution between the calculated similarity indices and the basket sizes of items for the cases.

As expected, there is an inverse correlation between the size of the suppliers' baskets and their similarity, meaning that the highest similarity indices between suppliers occurred in cases where the size of their baskets was a very small proportion of the total. Similarly, suppliers with the largest baskets of items (up to 60%) exhibited the lowest Jaccard similarity indices (below 23.23%).

It is important to highlight that there is a small group of suppliers that had 100% Jaccard similarity. This means that they bid for exactly the same items in all the tenders they participated in within the database universe. Although these cases are suspicious, this alone is not conclusive evidence of simulating non-existent competition. However, the intersection of these cases with those in the previous indicators may suggest the need for them to be prioritised in a detailed analysis and investigation.

5. Final considerations

This paper presented a case study with the bidding registrations of a mixed-economy company in the format of a graph. With the graph in hand, three indicators were developed to show that companies may have repeatedly coordinated their actions. The results obtained demonstrate the effectiveness of the graph model for detecting anti-competitive behaviour in tenders, as it facilitated the discovery and visualisation of hidden and indirect connections within the data, as well as discriminating a low proportion of cases against a large existing universe.

From a legal perspective, the detected cases represent signs of fraud but do not prove the intent or fraudulent nature of the act. To prove this, an internal investigative process is required, which should include hearings and the collection of documentation to support the arguments. However, this effort was not within the scope of this study, which was limited to the analytical effort of the data recorded in the databases. Nevertheless, the suspicious cases detected by this study provide elements that can support future investigations or internal audits.

Finally, the techniques and approaches carried out in this study can contribute to the improvement of systems to combat fraud in bidding processes and encourage graph modelling for other problems related to fraud detection. As possible future developments of this study, the experimentation with new algorithms and graph measures to detect other suspicious patterns is envisioned.

References

ACFE (Association of Certified Fraud Examiners). (2024) Occupational Fraud 2024: A Report to the Nations. https://legacy.acfe.com/report-to-the-nations/2024/

Amaral, W.S. (2020). Análise de Grafos para Apoio em Auditoria de Licitações Públicas. https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/trabalhoConclusao/viewTrabalhoConclusao.jsf?popup=true&id_trabalho=103180471-12.

Brazil (2024). Painel de compras. https://paineldecompras.economia.gov.br/

Costa, C. C. D. M., & Terra, A. C. P. (2019) Compras públicas: para além da economicidade.

Liu, J., Bier, E., Wilson, A., Guerra-Gomez, J. A., Honda, T., Sricharan, K., Gilpin, L., & Davies, D. (2016) Graph Analysis for Detecting Fraud, Waste, and Abuse in Healthcare Data. AI Magazine, 37(2), p. 33-46.

OECD (Organisation for Economic Co-operation and Development). (2021) Combate a cartéis em licitações no Brasil: Uma revisão das Compras Públicas Federais. https://www.oecd.org/competition/fighting-bid-rigging-in-brazil-a-review-of-federal-public-procurement-pt.htm

Padhi, S. S. and Mohapatra, P. K. J. (2011) Detection of collusion in government procurement auctions. Journal of Purchasing and Supply Management, 17(4), p. 207–221.

Paulo, A.C.R.M. and Rodrigues, A.P.S.P. (2022) Visualização de relacionamentos utilizando grafos como ferramenta de fiscalização de recursos públicos. Academic Journal on Computing, Engineering and Applied Mathematics, 3(2), p. 1-12.

Rodrigues, V. F., Policarpo, L. M., da Silveira, D. E., da Rosa Righi, R., da Costa, C. A., Barbosa, J. L. V., Antunes, R. S., Scorsatto, R. and Arcot, T. (2022) Fraud detection and prevention in e-commerce: A systematic literature review. Electronic Commerce Research and Applications, 56, 101207.

da Silva, L. C., Junior, R. V. C., Lopes, H. A. and dos Santos, M. (2020) Utilização de técnicas de Mineração de Dados para detectar possíveis relacionamentos entre empresas participantes de licitações nas Forças Armadas." Acanto em Revista 7(7), p. 85-85.

Victor, A. O., Sales, L. A. M., Moreira, R. S., de Paula, T. S., de Moraes, C. E. C., Meirelles, T. P., Lima, L. G. G. B., Rocha, J. F. C. and Contursi, B. S. N. (2024) Detecção de Indícios de Fraude em Licitações através de Técnicas de Análise de Redes Sociais. In: Rio Oil & Gas Expo and Conference (in press).