Brazilian Government Procurements: an Approach to Find Fraud Traces in Companies Relationships

Rebeca A. Baldomir[1,2], **Gustavo C. G. Van Erven**[1,2], **Célia Ghedini Ralha**[1]

[1]University of Brasília - Computer Science Department
Brasília, Federal District, Brazil

[2]Ministry of Transparency and General Comptroller of the Union
Brasília, Federal District, Brazil

(rebeca.baldomir, gustavo.erven)@cgu.gov.br, ghedini@unb.br

Abstract. *Data mining has been an area of high visibility in recent years and many researches have shown good efficiency in this area to find information in large databases. This paper presents an approach to find fraud traces applying data mining techniques to public databases of the Brazilian Federal Government bidings. The aim is to find evidence of fraud, such as stunts and cartels. The task of finding fraud evidences in large amount of data is complex for auditors since they have correlate data. The proposed approach was used to develop a prototype which has been used by auditors in the Ministry of Transparency and General Comptroller of the Union (CGU).*

1. Introduction

In the context of procurement processes, Government Agencies use to order goods and services through some processes that might ensure a good (or the best) quality by fair price. In Brazil, there are several kind of the procurement process and they are often controlled by an information system called SIASG – Integrated System for General Services Management in the Federal Government.

In the set of procurement processes, one of the most important is called *Pregão* [1], a reverse auction process. In this process, the suppliers present their initial proposals and then can start to give bids to lower the prices until the lowest offer remain and the other competitor give up.

Although this process encourages competition between companies, they can use some mechanisms to fraud it [2]. Companies can associate with each other and make collusion and cartels, that is, they will agree which one will win each bid item before the date of the auction and thus they can continue to alternate between them for the best contracts in each process.

Another unfair and common technique occurs when an individual, or a group of partners, own multiple companies from the same segment or industry. In this case, the companies are used to simulate competition between them, although the winner will always be a part of the same group of partners. [3]

In case of collusion, fake companies or another type of fraud, the processes will be harmfull, since the price can be manipulated to a very high value, or the products may came from a low quality brand.

A useful technique for finding these groups or cartels is called association rules. This technique makes it possible to search for relationships between items in a transaction [4]. In our context, items can be understood as a transaction and the companies are products of these transactions. Companies that usually show up together have more chance to be related as a collusion.

This paper proposes an information system approach to search for possible irregularities in processes associated with a specific company, in order to improve the work performed by CGU auditors. The Apriori scope is pruned in database and the rules can execute online, improving the auditor's target analysis, and other systems can be integrated to make the rule more meaningful as well.

The rest of the paper is organized as follows: in Section 2 we present a short overview of background; in Section 3 we present related work; in Section 4 we present the developed prototype to search for collusion and the results gathered inside the Ministry of Transparency and General Comptroller of the Union[1] (CGU). Finally, the Section 5 presents conclusion and future work.

2. Background

This section presents the background required for a better contextualization of this work, including association rules, the Apriori algorithm and the application domain in the context of the procurement processes of the Brazilian Federal Government.

Association Rules

Association Rules is a data mining technique that aims to discover strong relationships between data, finding patterns in associated data. There are two important metrics that show whether the relationship found by the association rule is strong:

1. Confidence: represents the percentage of times that the rule appeared correctly. The higher the confidence, better is the quality of the rule.
2. Support: represents the chance of the rule appearing in the analyzed dataset.

Another very important variable in the association rules is *lift*, an affinity measure that tells whether the occurrence of itemset A is independent of the occurrence of itemset B if P($A \cup B$) = P (A) P (B) . Otherwise, itemset A and B are dependent and correlated. This definition can easily be adjusted to more than two itemsets. The relation between the occurrence of A and B can be measured by the formula:

$$lift(A, B) = \frac{P(A \cup B)}{P(A).P(B)} \tag{1}$$

If the result of the equation is less than 1, the itemsets A and B are negatively correlated, in other words, an occurrence of one leads to an absence of the other. However, if the result is greater than 1, they are positively correlated, which means that the occurrence of one side leads to the occurrence of the other. If the value is exactly 1, the values are independent and there is no relationship between them [5].

[1] http://www.cgu.gov.br

Table 1. Transactions used to exemplify the Apriori algorithm

	Item 1	Item 2	Item 3
Transaction 1	1	1	0
Transaction 2	0	1	1
Transaction 3	0	1	0
Transaction 4	1	1	1
Transaction 5	1	0	0
Transaction 6	0	1	0
Transaction 7	1	0	0
Transaction 8	1	1	0
Transaction 9	1	1	0
Transaction 10	0	0	0

Apriori Algorithm

In order to find strong associations between data, the Apriori algorithm was proposed [6]. Apriori is an algorithm that extracts high trust rules finding relationships between the data. This algorithm uses an iterative approach consisting of k-itemsets to be used to explore (k+1)-itemsets. For each set it is necessary to check the presence of each item and collect the items that satisfy the minimum support. The resulting set is denoted by L_1. Then L_1 is used to find L_2 which is used to find L_3, and so on, until no more frequent k-itemsets are found. The discovery of each L_k requires a complete check of the database.

Table 1 illustrates 10 transactions that will be used to exemplify the Apriori algorithm. Initially, sets are created with each of the transaction items and their support calculated (Table 2).

The next set will be consist of creating all possible combinations of sets containing 2 transactions, and discarding the ones with a support value below the desired one represented by red lines in Table 2. Using a 20% support, all sets will be combined generating new sets. This process is repeated until there are no more sets to combine.

Table 2. Support of each set using Apriori algorithm

Iteration 1	
Set	Support
1	0,6
2	0,7
3	0,2
Iteration 2	
1,2	0,4
1,3	0,1
2,3	0,2

Procurement Processes

Procurement processes are the way the Government purchases products or services. It is the form that ensures full competition among participants seeking to ensure

compliance with the constitutional principle of isonomy, where the selection of the most advantageous proposal for the administration is ensured [7]. There are several procurement processes modalities, with the *Pregão* being the most known one [1]. Procurement processes are a major target for corruption due to being linked to the financial system, besides containing flaws in its process. Therefore, auditing has been an important process in the search and prevention of irregularities in brazilian procurement processes.

3. Related Work

Nowadays, in Brazil, corruption is a central subject due its economics and politics situation. According to the Organization for Economic Cooperation and Development (OECD) report[2] of 2016, public procurements are extremely vulnerable to corruption. Most of 50% of the foreign bribery cases aim public procurements contracts. Therefore, several works have been proposed addressing the corruption problem. Table 3 summarizes the papers found and described in this section.

Table 3. Related Work compared to the present work.

Paper	Association Rules	Other ML technique	Other approach	Online system	Domain
Present work	X			X	procurement
Ralha & Silva (2012) [4]	X	X			procurement
Nguyen et al. (2017) [8]		X			supply chain
Hein et al. (2014) [9]		X			economic sector
De Padua et al. (2016) [10]	X	X			transaction
Minović et al. (2014) [11]			X	X	procurement
Erven et al. (2018) [12]			X		procurement

Ralha and Silva [4] propose a solution for cartel detection using data from ComprasNet, a Brazilian procurements database, using data mining alongside multiagent systems. The AGMI (AGent-MIning tool) is composed by autonomous agents arranged in three layers that deliberate about the data mining techniques and parameters to discover knowledge about suspicious groups of companies. The system generates several association rules with scores built from the business domain to identify which them have high quality.

Data mining is also used in Nguyen et al. [8], where the state-of-art of data mining in supply chain management is presented and in Nelson Hein et al. [9], where factorial analysis is used to select social performance indexes of the most representative companies in the cyclic consumption sector.

[2]http://www.oecd.org/gov/ethics/Corruption-Public-Procurement-Brochure.pdf

De Padua et al. [10] uses community detection between transactions' items to identify clusters to split the database among them, and reach better rules, since several of them do not usually implies in a good relation. This because rules with support and confidence higher would be obvious.

Minović et al. [11] proposes introducing semantic technologies to procurement processes in Serbia to enable processing data by machines. In this model, the specialists would be able to define business rules related to suspicious situations being alerted when one of them occurs. Thus, a final application would previously recognize acquisitions that are potentially harmful and take actions to avoid them.

Erven et al. [12] presents a proposal of graph database model to identify relationships between companies that participate in Brazilian public procurement processes. Partners relationships and other data are inserted into a graph database and the shortest paths are found between companies that attend to the same acquisition process. A notation is proposed to model de database and it is compared against a relational database. At least, several queries is performed looking for insight the would be fraud evidences.

4. Association Rules in Procurement Processes

A prototype was developed in order to assist the CGU auditors to find fraud in the relationship between companies participating in procurement processes. To achieve this goal, association rules between data were applied through the Apriori algorithm.

4.1. Methodology

The model presented in Figure 1 shows how the data mining phases were used in this work. The first step is the Business Understanding where you will have adequate

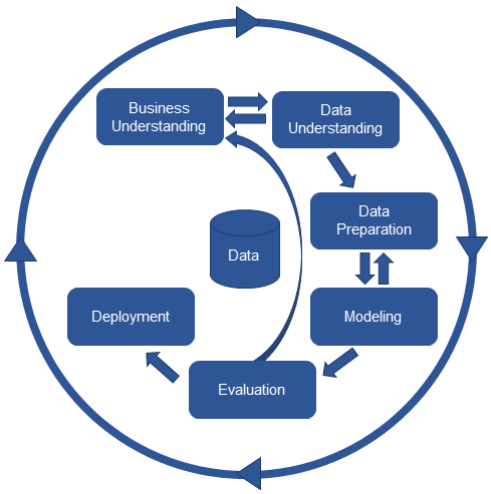

Figure 1. Data Mining Phases.

understanding of the problem that needs to be solved. The Data Understannding is the phase where you will understand the data and leave them in the correct way to use. This work will be focused on the next steps of the model that are Data Preparation, Modeling and Evaluation. The Deployment step was done at the CGU where auditors were able to use and evaluate the application.

Data Preparation

Data Preparation is the step where the interesting data was chosen from all available data in ComprasNet, a web site set up by *Ministério do Planejamento, Desenvolvimento e Gestão* [3] (MPOG) in order to provide information on the procurement processes and contracts promoted by the Federal Government to society. Of all existing data available, only 6 variables were considered relevant for this project: the purchase identifier, the CNPJ of the company that participated in the procurement process, the reference date, the value of the purchase and the modality of the procurement process. These columns were chosen because they are the only ones required to identify relationships between companies.

Modeling

Data mining was done using the pre-processed data already available in the database. The algorithm used in this work was Apriori which will be detailed in 4.2.

Evaluation

Having discovered the patterns of relationship between companies through the association rules, they will be presented to CGU auditors through a web interface. Therefore, the interpretation and evaluation of this information will be done by CGU's own auditors.

4.2. Prototype

The prototype was developed using two programming languages, Python[4] and R[5]. This decision was made in order to integrate this tool with other tools already deployed in CGU which were also developed in Python. It would facilitate integration if this prototype was also developed in Python. The R language was used for running the Apriori algorithm, using an existing R package. In order to integrate both languages, the rpy2 library was used, enabling the execution of R code from the Python application. The prototype was subdivided into three main components: Web, Data and Apriori according to the Figure 2.

Web

The Web component is responsible for receiving a certain CNPJ through Hypertext Transfer Protocol (HTTP) requests and sending this CNPJ to be processed. Once the processing is finished, this component shows the results in the Hypertext Markup Language (HTML) format.

This component was developed using the Python language. CherryPy [6] was chosen to build it, as it is a framework that allows the creation of web applications in the same way as any other object-oriented Python program. This results in smaller source code developed in less time.

[3] http://www.planejamento.gov.br
[4] https://www.python.org/
[5] https://www.r-project.org/
[6] http://cherrypy.org/

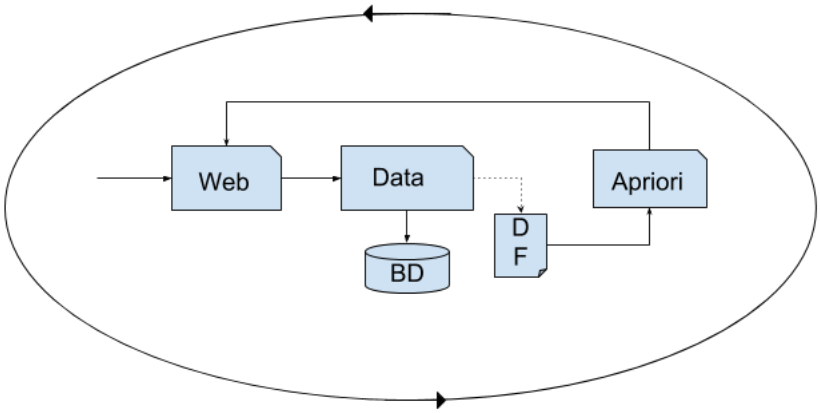

Figure 2. Prototype components

Data

This component is responsible for accessing the database and returning a list with CNPJs that participated in a purchase alongside the informed CNPJ in the Web component. The component makes this list available to the Apriori component through a data structure called dataframe.

The Data component was also developed in Python. The database accessed by this component was modeled using the MySQL Workbench tool Community [7] in version 6.3.9. The created database table contains a column for each column present in the data extracted on the pre-processing step.

Apriori

Association rules were applied in the procurement data available at ComprasNet in order to identify the possible correlations between the companies.

The algorithm used to apply these association rules was Apriori, which recognizes frequent items. For an item to be considered frequent, it must appear in the transactions in at least an amount equal to the minimum support. The set of items analyzed by the algorithm will be the CNPJs returned by the Data component.

The Apriori algorithm was applied using an implementation in the R language.

To exemplify the application of Apriori in the procurement process data, CNPJ 007XXXXXXXXXX was used. One of the rules resulting from the application was the following:

```
                [1]
039XXXXXXXXXX
                →
676XXXXXXXXXX
            0.3333333
                1
```

[7] https://www.mysql.com/products/workbench/

Notice that in this rule, specifically in Lines 1 and 3, that the CNPJs related by the rule are presented. In Line 4, the support is presented and in Line 5, the confidence used in the application of the rule is presented. It can be concluded from this rule that in 33 % of the purchases made between CNPJs 007XXXXXXXXXX and 039XXXXXXXXXX CNPJ 676XXXXXXXXXX was also contained in the set of buyers.

4.3. Results

The prototype's interface is shown in Figure 3. Note that it contains the CNPJ of the companies participating in the rule, the number of victories of each company participating in the procurement process, the lift, confidence and support used in the Apriori algorithm.

The search field can be used to search for the CNPJ present on the left side of the rules generated by the algorithm. The column *Company* and the column *Participated with* are the companies that participated in the procurement process with the CNPJ searched and appear on the right side of the rules. There are no limits regarding the number of companies that can participate in the bids together with the CNPJ searched.

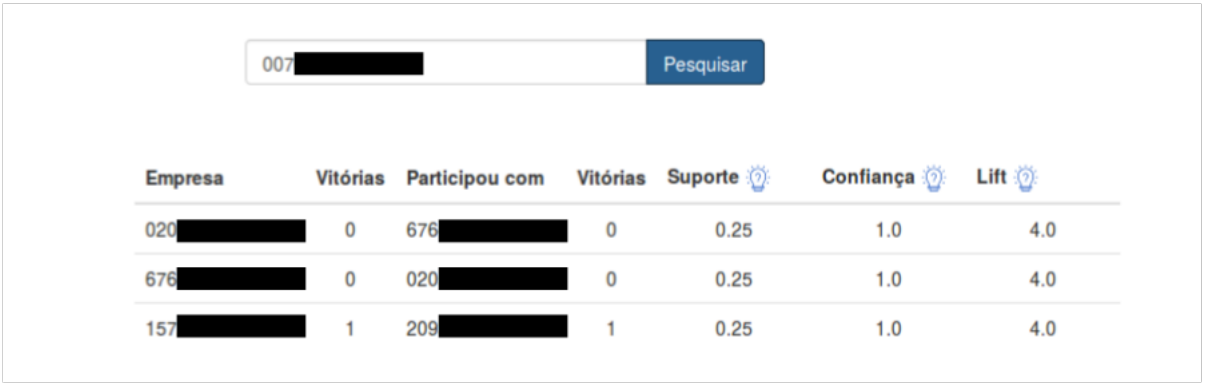

Figure 3. Prototype interface.

Initially, the prototype was executed locally with access to a base of 4,482,006 records (Table 4). Afterwards, the first tests in CGU were executed in order to collect feedbacks of the prototype's execution. Today, the developed prototype is executing in CGU with access to a base of 123,940,403 records that correspond to 20 years of procurement process data from 1997 to 2017.

Table 4. Number of records in relation to the execution environment

Execution Enviroment	Number of Records
Local	4.482.006
First tests at CGU	37.522.993
Real enviroment at CGU	123.940.403

An interesting use case found for the prototype was to use it in conjunction with other existing tools on CGU. Particularly, the use of the prototype alongside Yggdrasil, a CGU tool that searches for links or relationships between CNPJs, made it possible to

find a link between associates from two different companies who were participating in a procurement process together, which could be an indicative of fraud. The prototype built in this study was used to filter possible CNPJs in order to limit the search space, given that there are about 500 million CNPJs that could have a link. This way, the interesting CNPJs were narrowed down to those who appeared together in an association rule produced by the prototype, which made the search much easier.

Figure 4 illustrates the link found by Yggdrasil. The circles in red are companies' CNPJs and the link is represented by an edge between two vertices on the graph.

Figure 4. Link between two companies found in Yggdrasil system.

5. Conclusion

The CGU has a responsibility to fight corruption and for that, it is necessary to consider frauds such as stunts and cartels. This work proposes the use of data mining in procurement process data in order to find relationships between the participating companies.

In order to discover the relationships between companies in the Federal Government bidding processes, CGU's auditors would have to manually analyze all procurement process data available on the ComprasNet, which makes this task unfeasible. Thus, this work was developed in order to assist the auditors work on the discovery of relationships between companies participating in procurement processes.

The auditor familia with such relationships can check if there are other indications of a fraudulent relationship using another system already built by CGU that verifies links between companies.

The prototype developed during this study allows for the scope of analysis to be broadened regarding a certain CNPJ, and can be complemented by the analysis of links between companies. Finding all the links, or smaller paths, coupled with a base of over 500 million relationships would be computationally costly and only a few relationships

would really be relevant. However, in conjunction with the prototype, it may indicate which sets of companies are most interesting to analyse.

Acknowledgment

The authors would like to thank the CGU for support. C. G. Ralha would like to thank the Brazilian National Council for Scientific and Technological Development (CNPq) for the research productivity grant in the Computer Science area (PQ-2), process number 303863/2015-3.

References

[1] L. A. Joia and F. Zamot, "Internet-based reverse auctions by the brazilian government," *The Electronic Journal of Information Systems in Developing Countries*, vol. 9, no. 1, pp. 1–12, 2002.

[2] S. Rose-Ackerman and B. J. Palifka, *Corruption and government: Causes, consequences, and reform*. Cambridge university press, 2016.

[3] G. C. van Erven, M. Holanda, and R. N. Carvalho, "Detecting evidence of fraud in the brazilian government using graph databases," in *World Conference on Information Systems and Technologies*. Springer, 2017, pp. 464–473.

[4] C. G. Ralha and C. V. S. Silva, "A multi-agent data mining system for cartel detection in brazilian government procurement," *Expert Systems with Applications*, vol. 39, no. 14, pp. 11 642–11 656, 2012.

[5] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, Inc, 2005.

[6] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*, ser. VLDB '94. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, pp. 487–499.

[7] Brasil, "Lei nº 8.666, de 21 de junho de 1993," Jun. 1993, disponível em: $http://www.planalto.gov.br/ccivil_03/leis/L8666cons.html$. Acessado em: 07/11/2017.

[8] T. Nguyen, Z. Li, V. Spiegler, P. Ieromonachou, and Y. Lin, "Big data analytics in supply chain management: A state-of-the-art literature review," *Computers & Operations Research*, 2017.

[9] N. Hein and F. Kreuzber, "Aplicação da análise fatorial como ferramenta de data mining no desempenho social das empresas do setor de consumo cíclico listadas na bmf-bovespa," 2014.

[10] R. de Padua, E. L. S. Junior, L. P. do Carmo, V. O. de Carvalho, and S. O. Rezende, "Pre-processing data sets for association rules using community detection and clustering: a comparative study."

[11] M. Miroslav, M. Miloš, Š. Velimir, D. Božo, and L. Đorđe, "Semantic technologies on the mission: Preventing corruption in public procurement," *Computers in industry*, vol. 65, no. 5, pp. 878–890, 2014.

[12] G. Van Erven, W. Silva, R. Carvalho, and M. Holanda, “Graphed: A graph description diagram for graph databases,” in *World Conference on Information Systems and Technologies*. Springer, 2018, pp. 1141–1151.