NOVA
IMS
Information
Management
School

MDSAA

Master's Degree Program in
Data Science and Advanced Analytics

Mapping a Co-competition Network of the Portuguese Public Procurement Market

Gonçalo Baptista Perdigão Silva

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Mapping a Co-competition Network of the Portuguese Public Procurement Market

by

Gonçalo Baptista Perdigão Silva

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science.

Supervised by

Advisor: Flávio Luis Portas Pinheiro, PhD, NOVA Information Management School

Co-advisor: Bruno Miguel Pinto Damásio, PhD, NOVA Information Management School

November, 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Amadora, 5/11/2023

DEDICATION

I dedicate this dissertation to my late grandfather, who I know would be very proud of this achievement in my life.

Thank you for everything.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my advisor and co-advisor, Prof. Flávio Pinheiro and Prof. Bruno Damásio, for the support and feedback provided during the development of this work. Additionally, I want to thank Niclas Sturm for the insights and feedback provided during our meeting.

I would also like to express my great gratitude towards my parents and girlfriend, whose encouragement, love, and moral support was essential in providing motivation to finish the project.

Finally, I want to acknowledge my friends and every person who has helped me during the entirety of my bachelor's and master’s degree in this college. Thank you for aiding me get where I am today and achieving this goal in my academic pursuit.

ABSTRACT

Fraud in the public procurement sector is a persistent and challenging issue which can have harmful consequences in society, hindering economic and social growth, and corroding public trust in the government. This thesis aims to use network science methods to analyze and detect fraudulent activities by utilizing a Portuguese data set composed of thousands of public contracts between 2009 to 2023. The approach consists of utilizing the firms' co-occurring similarities to build a co-competition network and to detect communities of firms which are likely competitors and collaborators. With the aid of additional indicators such as activity diversity, regional diversity and single bidders, the characterization of the communities allows to emphasize groups of firms which are more probably engaging in illegal schemes such as collusion or market distortion. Furthermore, the work reinforces the efficiency of a network-based approach to unveil complex relationships, and in detecting and mitigating fraud. Finally, the findings and insights revealed in this study aim to aid in the development of fraud detection techniques and to provide a better understanding of the Portuguese public procurement environment.

KEYWORDS

Public Procurement; Network Analysis; Co-competition Network; Portugal; Clustering; Corruption; Community Detection

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

CPV Common Procurement Vocabulary

1. INTRODUCTION

The use of public procurement can effectively advance social, technological, economic, and environmental objectives (Browne, 2017). The amount spent on procurement has increased in recent years, and with it has come the public's demand for greater efficiency and openness (Browne, 2017).

To identify, investigate, prosecute, fight, and prevent public fraud is a challenge in most countries around the world (Gottschalk & Stanislas, 2017). Public procurement fraud and corruption have long been acknowledged as significant global challenges, costing governments a significant amount of money, weakening public confidence in the distribution of public funds, limiting investment and growth and leading to an ineffective government (Camões & Mendes, 2005; Rose-Ackerman, 1999).

According to OECD (2023) the total volume of public procurement accounted for 13% Gross Domestic Product (GDP) in OECD countries and its global expenditure is estimated at nearly 9.5 trillion US dollars (World Bank, 2020). The complexity of the process, as well as the many forms corruption it can take - embezzlement, undue influence in the needs assessment, bribery of public officials, fraud in bid evaluations, invoices, or contract obligations – makes this sector one of the most vulnerable to corruption (OECD, 2016).

The combination of high vulnerability and government expenditure dictates the high cost of corruption in public procurement. Measuring direct costs of corruption is challenging due to its covert nature, but it is estimated that corruption frequently results in losses of 20–30% of project value (Wells, 2013; Stansbury, 2005). In addition, is it known that more than half of foreign bribery cases occurred to obtain a public procurement contract (OECD, 2014) and more than 3 out of 10 companies that have participated in a public tender say corruption prevented them from winning (OECD, n.d.).

This level of threat and many other factors are the reason why it is crucial to fight these illicit activities and why, for example, the FBI has stated public corruption is its top criminal priority (Comey, 2016). Organizations like the World Bank, the United Nations, the OECD, the Council of Europe, and the European Union have all made joint efforts to combat corruption in the public sector. Several national legal systems contain institutions and legislation focused at preventing and prosecuting corruption in addition to international efforts to do so (Williams-

Elegbe, 2012). Combating corruption and improving public procurement can yield big savings: even a 1% efficiency gain could save €20 billion per year (DG GROW, n.d.).

Having that said, it is key to conduct scientific research and investigating cutting-edge approaches to detect fraudulent behaviors from firms in bidding markets (Lyra et. al. 2022). Network analysis can be a useful tool to unveil patterns, understanding relationships and detecting inconsistencies (Lyra et. al. 2022). More specifically, a network built on firm-firm interactions can highlight the existence of groups of firms where there is higher chance of exclusive interaction between themselves and collusion, making it possible to participate in bid-rigging, a typical kind of anti-competitive behavior in these marketplaces (Wachs & Kertész, 2019; Lyra et. al. 2021).

Lyra et. al (2021) look to identify groups of firms from the state of Ceará municipalities susceptible to forms of corruption from their co-bidding patterns as well as other characteristics, such as bidding coordination and diversity of activities. We embark on a similar path by exploring the background of public procurement fraud within the Portuguese environment, extracting insights from a dataset of public procurement contracts from 2008 to 2023. It looks to use network science methods to describe the connections between rival firms and find communities of firms which regularly compete in tenders with the same scope of firms, thus depicting a co-competition network. Furthermore, it will focus on exploring other attributes of the communities (Lyra et. al. 2021) which can be correlated to corruption. Ultimately, it will lean on a comparison and discussion of results of fraud in public procurement markets between Brazil and Portugal.

The next chapter shall aim to provide relevant literature review to the context and how this article can bring a new contribution to the literature. Next follows a detailed description of the data, its processing, and the methodology used to get the results. Following is the presentation, discussion, and analysis of results. Eventually, some final regards about limitations and future work.

2. LITERATURE REVIEW

2.1. What is Public Procurement?

Public procurement's goal is to quickly and economically award contracts to capable suppliers, service providers, and contractors for the supply of products, labor, and services to support government and public sector operations (Lynch & Angel, 2013). It refers to the government's acquisition of goods, works or services that are acquired from another party outside the procuring entity itself - often from the private sector, although it may be from another public body (Arrowsmith et. al. 2011). The concept of public procurement can be used to refer to three phases (Arrowsmith et. al. 2011): i) Decision of what goods or services are to be bought; ii) Placing a contract to acquire those goods or services, which involves choosing the contracting partner and defining the terms; iii) Administering of the contract to ensure effectiveness.

Growing interest in leveraging public procurement to promote innovation and development has been shown in recent years. In the European Union member countries, public procurement often accounts for 10–15% of a nation's GDP (EC, 2018). A growing number of governments argue that public procurement should be utilized more frequently and explicitly to support innovation, technology, and economic development (Lember et. al. 2013).

Understandably, being one of the essential economic activities of the government and representing a substantial amount of its GDP, it can largely contribute to the enormous swings in the economy's fortunes (Siyal & Xin, 2020). However, even though governmental organizations have eagerly sought to enhance public procurement methods, public procurement has been a field of research that has received little attention (Thai, 2001).

2.2. Public Procurement Corruption

Public procurement has been recognized as the key and essential tool to ensure that the administration of public resources is competent and free of corruption (Siyal & Xin, 2020). Despite this, it is known to be one of the public activities most vulnerable to corruption (IMF, 2019). There are several indicators during the procurement process which can be associated with corruption risk, such as single bidders, non-open procedures, lack of transparency, extensive or short bidding periods and spending concentration (Abdou et. al. 2022).

Furthermore, the high costs and uncertainty involved with public contract assignments serve as a catalyst for corruption (Søreide, 2002). To increase their chances of receiving government contracts, businesses may turn to corruption, ultimately leading to obstruction of markets and slowing down economic growth (Søreide, 2002). Indeed, there seems to be a particularly high potential for corruption in the public procurement sector (OECD, 2005), partially associated with the complexity of the procurement system, and the low-risk high-reward linked with fraudulent activities within the sector (LaCascia & Kramer, 2023).

Truly, public procurement corruption is a serious problem which can have very damaging consequences. What makes it even more threatening is the fact that it is the hardest type of fraud to detect. This makes it one of the biggest challenges in the public sector, being responsible for the loss of billions of dollars annually (CIPFA, 2020; OECD, 2005). Understandably, there is a need to build countermeasures and strengthen the public sector's defenses against corruption. For example, Magakwe (2022) has suggested that each country should build a strong legal framework based on existing guidelines provided by international organizations to enhance compliance with procurement processes. On the other hand, the use of technology and innovative practices should be used to fight, prevent, and detect public procurement corruption. That said, there has been a lot of work which has gone into developing analytical strategies to comprehend and reduce the effects of corruption (Fazekas et al. 2018). Particularly, the work of this article will delve into the use of network analysis for detecting corruption.

2.3. Public Procurement in Portugal

Portugal's expenditure on public procurement has been experiencing a noticeable growth over the last few years, representing more than 6% of GDP in 2021 – double the value of 2017 - and totaling 13.74 billion euros (IMPIC, 2021).

The Portuguese public procurement system is now centralized, presenting meaningful transparency policies and national implementation of an e-procurement system that has led to many positive impacts (EC, 2022). Despite the advancements in the public procurement process, Portugal still exhibits a high risk for corruption - lasting issues in the sector, such as bid rigging and conflict of interest are the result of poor monitoring of the procurement process (EC, 2022). According to Transparency International (2021), Portugal is one of the countries in Europe where the corruption level is most rapidly increasing, and where connection rate (percentage of users who used personal connections to get a service) presents one of the highest values amongst European countries; additionally, most Portuguese citizens think companies are not playing by the rules and that government corruption is an issue.

For additional context, Portugal has two autonomous administrative regions, the Azores and Madeira, which have independently customized regional public procurement laws to fit the unique needs of their respective areas.

2.4. NETWORK ANALYSIS FOR FRAUD DETECTION

As fraudulent activities become more developed, traditional methods of detection may have encountered limitations to keep up with these more sophisticated practices. Hence, network analysis is becoming an up-and-coming method to fight against fraud. Networks can provide the ideal framework for studying the organization of corruption (Wachs et. al. 2021). These are structures that represent relationships (edges) between objects (nodes). In the context of this work, nodes would be firms and the edges would represent their connections in the procurement process.

The work of Lyra et. al. (2022) identifies the most recent network science techniques which can be used for fraud detection in the procurement sector. Specifically, it highlights community detection as well as centrality measures as the most broadly used tools in scientific articles of this nature. Community detection is a widely used tool for cluster identification, mainly cartels. In a public auction market, cartels are groups of firms which may raise profits by establishing prices collectively (Marshall & Marx, 2014). The research made by Wachs & Kertész (2019) uses firms co-bidding behavior to generate a network of firm-firm interactions. These patterns of interactions measure collective behavior which can be effective for cartel identification. Although the approach is only suggestive, they find that collusion is more likely to occur among groups of firms where interactions are more cohesive and exclusive. In another research, Cheng et al. (2017) also applied community detection methods to a complex network to analyze bid rigging behavior and collusion. The network was built using the companies' co-bidding frequency values (i.e., number of times companies bid in the same tender) as weights. The results show that communities with densely connected nodes that present large weights indicate cooperative behavior among enterprises within such communities. Similarly, Zhu et. al. (2020) built a network utilizing the co-participation rate of companies in the same tender, followed by identification of communities using a modularity optimization algorithm. Additionally, they detect network overlaps to further divide communities into new ones. The results show that communities with overlapping nodes presenting high degree and centrality indicate a significant level of involvement in collusion. On another note, Grassi et al. (2019) gathered evidence that betweenness centrality can be effective on identifying mafia leaders within a criminal network. They show that nodes with high centrality measures are significantly more likely to be leaders with strong brokerage positions, thus having privileged access to information.

Wachs et al. (2021) characterize the distribution of corruption risk in EU countries by using bipartite networks mapped from procurement markets. They quantified single bidding across several communities in a network and found evidence that corruption is not randomly distributed in the public procurement sector. Additionally, they highlight that more instances of single bidding indicate a higher probability of finding a corrupt actor among the neighbors. Fazekas et. al. (2020) utilize a network framework to analyze the connection between corruption and market structure. They found that corrupt behavior in the procurement sector mainly revolves around favoring certain suppliers while purposely excluding others. Moreover, they claim that merely looking at networks can provide insight into the nature of buyer-supplier connections more than standard statistical analysis.

As far as research in a Portuguese environment, Carneiro et al. (2020) present a prototype system which utilizes a network combined with a rules engine and a decision support system for fraud detection. They argue that the system's facilitation of information access and compliance checks will lead to more transparency in public procurement.

In summary, the reviewed literature illustrates the important role of network analysis for fraud detection in various sectors, mainly the public sector. This approach can overcome limitations found in traditional statistical methods. Its significance lies in its capacity to reveal deep connections and interactions among entities involved in corrupt schemes. However, since network science methods do not always provide exact results, they are usually utilized together with other metrics such as single bidder, neighbor diversity, winning rate and diversity or even other indicators are for more accurate findings (Lyra et. al. 2022).

3. DATA & METHODOLOGIES

3.1. DATA

Data was retrieved from the web portal https://www.dados.gov.pt/ (accessed in March 2023), provided by *Instituto dos Mercados Públicos, do Imobiliário e da Construção* (IMPIC). The dataset consists of Portuguese public procurement contracts ranging from the year 2009 to 2023.

The initial dataset comprised of 1,557,278 rows, each one representing a different contract. It contained information - name and VAT number - about the company which issued the contract, the contracted company as well as the other contestants, and more relevant information such as contract value, location, contract type, CPV code, procedure type and date, among other non-essential attributes. It's important to note that some of the variables were stored as *json* arrays, containing information to create several other columns (see Table 1, step 6). The first step was to explore and clean the data in order to create a ready to work dataset. The data was processed as follows:

1.	Check and remove duplicates
2.	Consider only contracts with *Open Procedure* as procedure type and remove the rest of the rows. These are the relevant contracts to the purpose of this thesis.
3.	Handle missing values: All rows where essential information was missing were removed (e.g., Contracted firm, contract type, procedure type)
4.	Change datatypes
5.	Split and extract information: Contestants – extract VAT number and company name; Location of execution – extract municipality and district. New columns with this information were created while old ones were discarded.
6.	Creation of new binary column: a value of 1 if company was contracted, else 0.
7.	Check missing values of new columns: removed all rows where the contestant's VAT number was missing, resulting in the deletion of 2.6% of the dataset.
8.	Validate VAT numbers[1] and CPV codes and remove non-valid rows.
9.	Coherence checks. Remove: contracts with publication date earlier than signing date; negative contract values; impossible dates
10.	Remove all non-needed columns

Table 1 | **Pre-processing steps.**

It is worth mentioning that the newly created columns about the district and municipalities of execution presented more than 90,000 missing values. Since this represents a good portion of

[1] The VAT numbers were validated according to the Portuguese enterprise VAT number rules. Can be found at https://www.nif.pt/nif-das-empresas/

the data set and there is not a sensible method to fill these values, the decision was to keep them. These features will not be relevant to creating the network itself but will have an impact on later stages of the work.

The resulting dataset is composed of 736,638 observations, where each row represents a firm's bid to a tender, hence having a bipartite nature. At this point, the dataset includes 136,996 tenders and 17,676 distinct firms.

Hereafter, the dataset was divided into three new datasets:

I. ***tender_firm*** - Each row represents a firm's bid to a tender (bipartite dataset).

II. ***contract_data*** - Contains information about each tender. Each row is a different contract.

III. ***firm_data*** - Contains information about each firm, specifically how many contracts of each type of each firm has bid on. To obtain this, all contracts were grouped into the three main common types of contracts (Arrowsmith et. al. 2011) – Goods, Services and Works – and Others if contract type was not clear.

tender_firm	***contract_data***	***firm_data***
Contract ID	Contract ID	Firm VAT number
Firm VAT number	Contract Type	No. of Goods Contracts
CPV code	CPV code	No. of Services Contracts
Publication Year	Price	No. of Works Contracts
	Execution District	No. of Other Contracts
	Execution Municipality	
	Publication Year	
	Region	

Table 2 | **Data sets' variables.**

A new column - *Region* - was created, indicating the Portuguese region of the execution of the contract. It was built based on the district of the execution. The correspondence district-region can be found in the appendix (Table A) and is based on information provided by INE (2015). Naturally, since a lot of the district values were missing, there will be the same number of region values missing. Moreover, some district values contained multiple districts, each from a different region. Hence, the resulting value of the column *Region* will contain both regions of execution.

To better understand the data and extract useful and valuable information, a small statistical analysis was conducted. Figure 1 demonstrates the clear growing tendency of the number of public contracts in Portugal during the last few years.

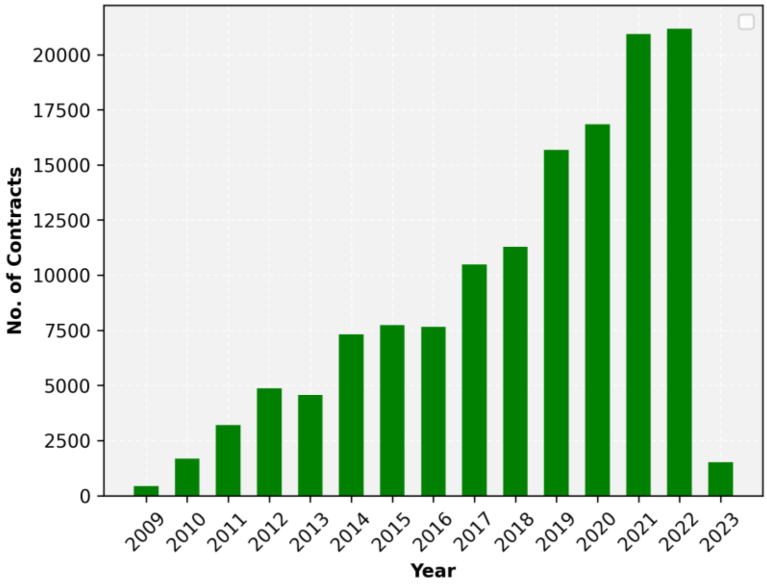


Figure 1 | **Yearly Contract Distribution**. Represented in this figure are the number of open procedure contracts per year. It can be observed that there has been a constant growth in the number of contracts during the past decade. **Note:** This data was obtained in March, 2023, hence the number of contracts in 2023 is so low.

The total contract amount per year, in euros (€), is displayed below in figure 2. Values per year mostly range between 1 to 10 billion (€). A growing tendency is also noticeable, which makes sense according to the growing number of contracts. Additionally, the average contract price is around 335 thousand euros, and the highest contract value is about 3,78 billion euros, from a contract of 2021. This contract alone explains the visible spike of total contract value in the year 2021 in figure 2.

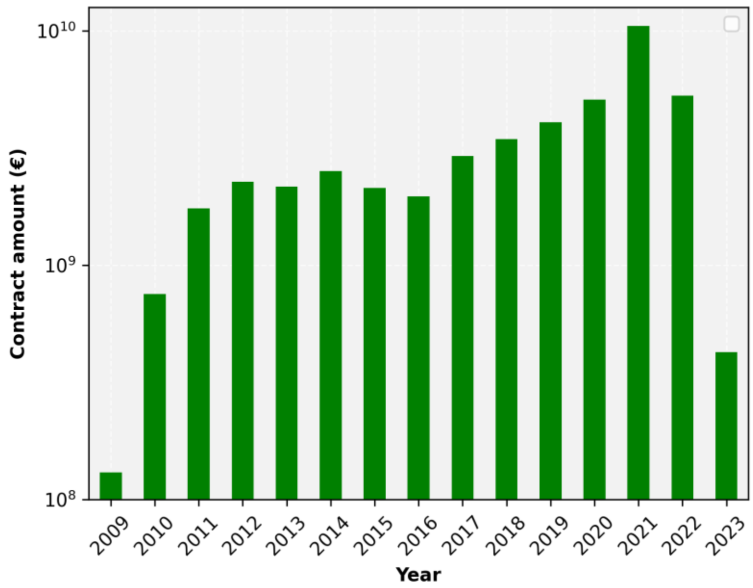


Figure 2 | **Contract amount (€) distribution**. This plot displays the total amount of public contract values grouped per year. **Note:** This data was obtained in March, 2023, hence the total amount of contract value in 2023 is so low.

Moreover, figure 3 shows the distribution of contract types. Considering the open procedure contracts, *Goods* seems to be the most prevalent contract type in Portugal, with more than double the amount of each of the other two types of contracts.

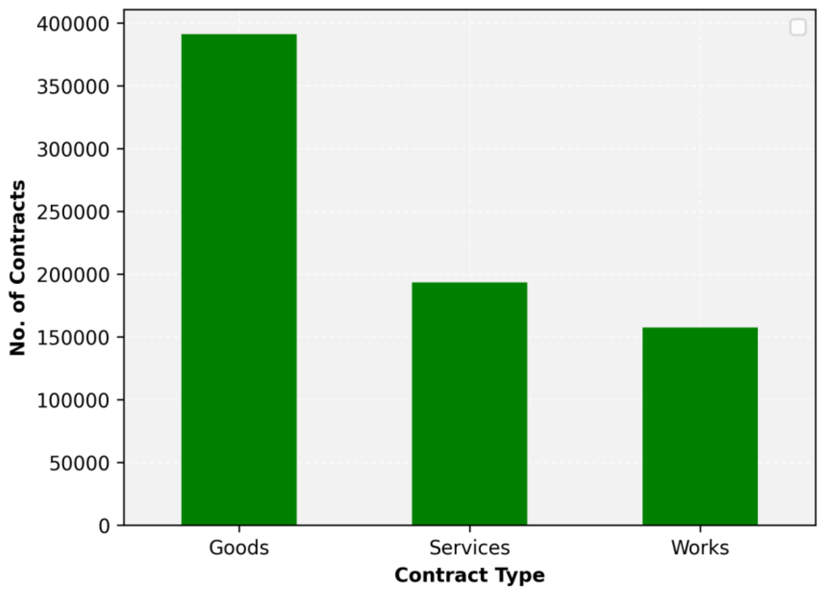


Figure 3 | **Contract type distribution**.

In figure 4 we observe that most firms have won either zero or a low number of contracts, and that a minority of firms have won a big portion of the contracts. This suggests that a small number of firms control the contract acquisition process. On average, each firm won 12.2 contracts, with the highest winning firm having 1277 contracts won.

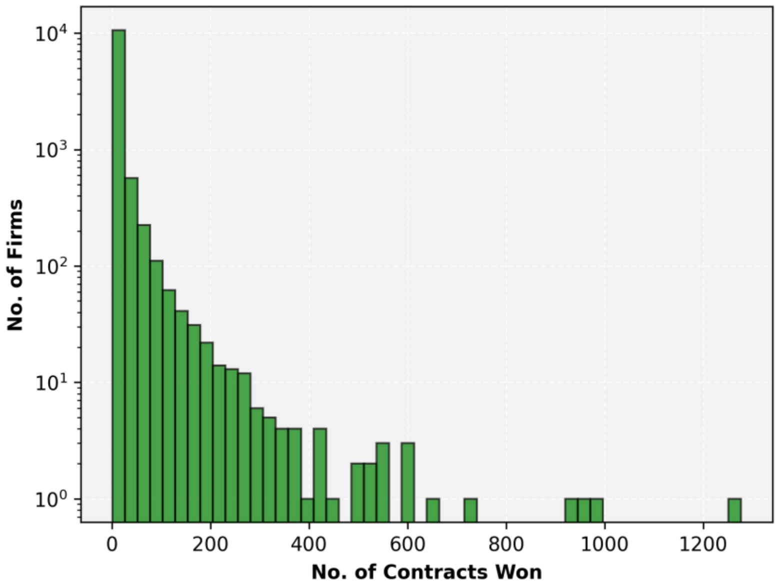


Figure 4 | **Distribution of contracts won by firm**.

3.2. COMMON PROCUREMENT VOCABULARY

The common procurement vocabulary (CPV) is a classification system that defines the subject of a public procurement contract. To describe various product and service categories, it employs a coding system made up of 9 characters (8 digits and a control character). These codes are built on a tree structure linked to a definition of the goods, services, or works that make up the contract. The first two digits (**XX**000000-Y) refer to one of the 45 divisions (e.g., Real estate services, Construction work, etc.), which are the first level of the hierarchical structure. Each digit after the first two adds another degree of classification and level of detail. An example is illustrated in Table 3.

In the context of this thesis, we are interested in grouping firms by type of contract. Since the CPV code gives more detailed information about the subject of the contract, we'll be using it to extract a bigger amount of detail. For that end, we trimmed down each CPV code to the first 2 digits, assigning each contract to one of the 45 divisions.

Level	Code	Subject
Division	**03**000000-1	Agricultural, farming, fishing, forestry and related products.
Group	**031**00000-2	Agricultural and horticultural products.
Class	**0311**0000-5	Crops, products of market gardening and horticulture.
Category	**03111**000-2	Seeds.
Sub-Category	**031111**00-3	Soya beans.

Table 3 | **CPV code example.** The example displays part of the CPV hierarchy to illustrate how a contract is classified.

3.3. Network Deduction

Since the goal is to study the relationships between firms using their co-occurring patterns, we need to focus on the construction of the firm-firm projection (Fierăscu, 2017). We use cosine similarity (Sidorov et. al. 2014) to uncover underlying patterns from the firms' co-occurrence matrix to build the network. The final network is interpreted as a co-competition network, representing collaborative or cooperative relationships between firms (Gulati et. al. 2012).

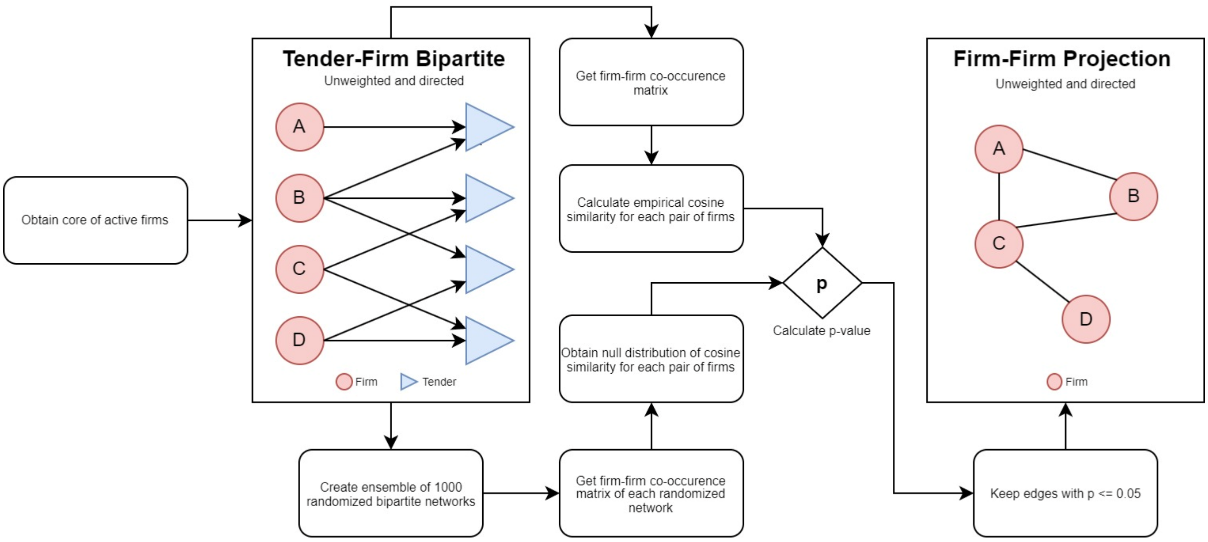


Figure 5 | **Firm-Firm projection inference steps**. **Note:** Adapted from Lyra et. al. (2021).

The first step was to obtain the core of active firms in the dataset. To do so, the average number of bids per firm was calculated - on average, each firm bid a little over 41 times. The decision was to keep the firms that bid more than 40 times, meaning that the remaining firms could be classified as moderately active in the time window of the dataset, at the very least. Figure 6 compares the original dataset (P_{all}) with the resulting filtered dataset ($P_{filtered}$). The main conclusion we can draw is that the distribution of the number of bidders per tender remains constant even after applying the filters. When observing the number of bids per firm, the OLS regression lines slopes are similar, suggesting consistent linear scaling relationships. However, the power-law scaling exponents of the number of bids per firm of both datasets are not so similar. This is to be expected since the applied filters basically remove all the extreme values from the upper tail, but we accept such consequence to gather the most active firms. Hence, the final working dataset, represented by a bipartite structure, includes a total of 633,848 bids made by 2809 firms to 122,974 tenders.

Next, we compute the co-occurrence matrix of the firm nodes. In this work, we leverage the use of the co-occurrence matrix, an important tool with several applications in information science which can be used to find relationships between the firms. The co-occurrence value between each pair of firms can be computed as:

$$C_{ij} = \sum_k (b_{ik} \times b_{jk}), \tag{1}$$

where b_{ik} is one if firm *i* made a bid to tender *k* and zero otherwise; b_{jk} is one if firm *j* made a bid to tender *k* and zero otherwise. The resulting matrix contains the number of times each pair of firms bid together on the same tender.

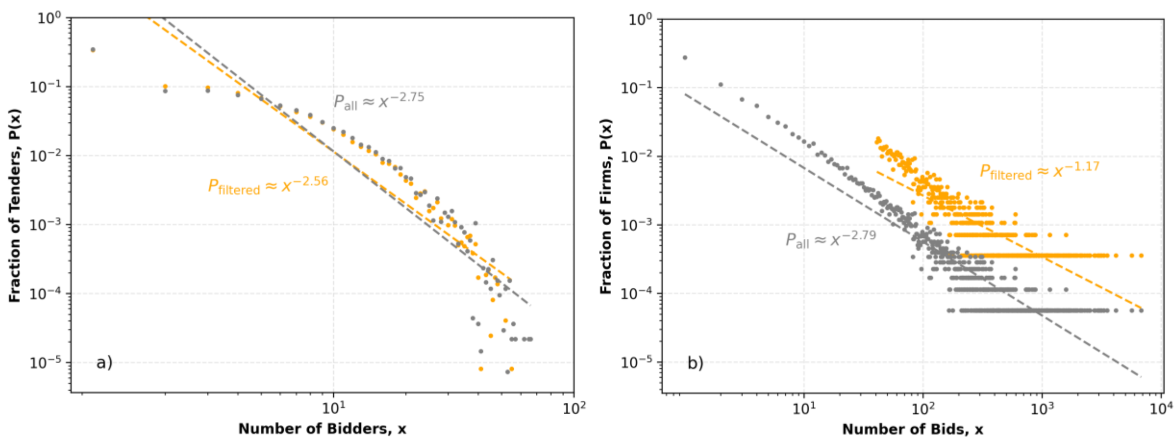


Figure 6 | **Filtered vs non-filtered dataset distributions.** Panel **a)** compares the frequency of bidders per tender in the original dataset (gray) with the filtered/working dataset (orange). Panel **b)** compares the frequency of bids per firm in the original dataset (gray) with the filtered/working dataset (orange). The dashed lines represent the OLS regression lines for each set, demonstrating the linear relationships. In addition, power law distribution expressions (P_{all} and $P_{filtered}$) are superimposed on the data to reveal potential scale-free patterns. Adapted from Lyra et. al. (2021).

We then estimate the cosine similarity between each pair of firms by applying it on the co-occurrence matrix (Zhou et. al. 2016). Cosine similarity is a frequently employed metric in data analysis, information retrieval and natural language processing. In this case, it is particularly valuable to compare co-occurrence patterns from the firm's bids. A value of *1* between a pair of firms would mean that each of those two firms participated in the same number of tenders with each firm from the same set of firms. On the other hand, a value of *-1* would mean that the two firms don't have a single firm in common with which they have bid on a tender with. Formally, cosine similarity between a pair of firms can be described as:

$$C_{ij}^{s} = \frac{S_{ij}}{\left(\sqrt{S_{ii}} \times \sqrt{S_{jj}}\right)}, \tag{2}$$

where S_{ij} represents the dot product of the co-occurrence vectors from firm *i* and *j*; $\sqrt{S_{ii}}$ and $\sqrt{S_{jj}}$ represent the Euclidean norms of the co-occurrence vectors from firm *i* and *j*, respectively. The resulting matrix is what we call empirical cosine similarity matrix, containing the real cosine similarity values between each pair of firms.

A projection is a technique used to project a set of nodes in a bipartite graph onto new a graph, with the goal of studying the relationships within that set. Each pair of nodes has a weight representing the number of common nodes from the other node set. In our case, the firms are the nodes to be projected and their pairwise cosine similarities are used as the weight. However, the use of a projection on a bipartite structure can result in loss of information which can be manifested by the creation of non-significant links (Piccolo et. al. 2018).

To mitigate this problem, we estimate the significance of the empirical C_{ij}^{s} to test the hypothesis that $C_{ij}^{s} > 0$. By creating an ensemble of 1000 randomizations of the initial bipartite network (Lyra et. al. 2021), we generate a null distribution of cosine similarity for each pair of firms. The aim is to compare the observed cosine similarities with the distribution obtained from the null model and ultimately test if they are significantly different from what the null model would predict by chance. Each randomization consists of: i) randomizing bipartite data - while ensuring that the number of firms bidding on each tender and number of bids per firm remain constant; ii) generating the co-occurrence matrix; iii) obtaining the cosine similarity matrix.

Finally, we calculate the one-tailed p-value associated with the observed empirical cosine similarity. It is acquired by determining the likelihood of obtaining a value from the cumulative distribution of values in the null distribution that is as, or more extreme than the observed empirical cosine similarity. Edges with a p-value greater than 0.05 as well as edges with negative cosine similarity were discarded.

The developed firm-firm co-competition network is composed of 2801 nodes and 130,471 edges. It presents an average degree of 93.16 (Newman, 2010) and a clustering coefficient of 0.74 (Newman, 2020; Girvan & Newman, 2002).

3.4. Community Detection

Given the current network, the final step to prepare for the analysis is to identify communities. Communities, or *clusters*, are essentially groups of cohesive nodes which are likely to share common characteristics or serve similar functions within the context of the network (Fortunato, 2010). Therefore, community detection is a crucial technique in the context of network analysis. It can help identifying the underlying structure of the network, as well as reveal patterns and relationships between groups of actors which may share interests or affiliations.

For that reason, we explore and identify communities with the use of *Louvain* algorithm (Blondel et. al. 2008). This method is known for being extremely fast, scalable and for optimizing the *modularity* (Newman, 2006) of the network. Simplified, it consists of two phases (Blondel et. al. 2008): i) each node is assigned to a community such that the modularity of the network is maximized; ii) Generating a new network whose nodes are now the communities found during the first phase. Finally, this process is iterated until a maximum of network *modularity* is achieved.

Hence, the method was applied to the firm-firm projection. Since we are using a network constructed on the firms' co-occurring similarities, the expectation is that firms in each community will tend to have more significant links or similarities in terms of the other firms they usually bid with. In other words, their co-competitors. The term *co-competitor* describes a situation where firms can simultaneously work as competitors and collaborators.

3.5. ACTIVITY DIVERSITY

As previously mentioned, we'll be looking to explore additional metrics that could give relevant information about the communities. For that end, we utilize the *Simpson's diversity index* (Simpson, 1949). It measures the probability that two randomly selected individuals from a community share common features. Accordingly, a lower value (closer to 0) would indicate high diversity whereas a higher value (closer to 1) would indicate lower diversity. The Simpson's index for each community can be estimated as (Lyra et. al. 2021):

$$D_S^{C_i} = \sum_{t \in \gamma} (p_{C_i}^t)^2 , \tag{3}$$

where $p_{C_i}^t$ corresponds to the proportion of bids made by a community $C_i, \forall_i \in [1,2, \ldots, n]$ where n is the number of communities, in a certain contract type γ: { Construction work, Financial and Insurance Services, Medical and Pharmaceutical, …} or Portuguese region γ: { Lisboa, Norte, Alentejo, …}. Each community's $p_{C_i}^t$ quantity is normalized so that $\sum_t p_{C_i}^t = 1$. As a result, in the context of this work, this index will measure the probability that two bids made by firms within the same community share contract type or region of execution.

4. RESULTS AND DISCUSSION

4.1. CO-COMPETITION NETWORK

The Louvain method identified 26 communities with a modularity of 0.75. Before progressing, we compute the number of contracts in each community per the previously obtained Portuguese regions and divisions of CPV codes. The region of execution and CPV division of contracts each community usually bids on will be their main identifiers. In other words, these will be the features that characterize each community.

Based on this new information, some communities with similar contract execution regions and CPV codes were joined to form a bigger community. Hence, the final the number of communities is 23 with a modularity of 0.74. Each community is labeled as *C1*, *C2 …C23* for ease of identification. The high value of modularity in the network indicates that communities are effectively distinct and distinguishable. This can be explained by the high specialization of each community in contract type and region of activity, but mainly the prior.

To smooth the visualization, Figure 7 presents the graphical representation of the 15 biggest communities in the network - community sizes can be found in Table B in the appendix. The number of bids per CPV code and region were computed for every community. Each community is represented by a different color and their name is associated with the most common type of contract within itself, as well as the region of execution. If the name does not contain the region, it is because there was not a region which stood out enough from the rest.

The graph shows a main component in the center with several disconnected components in the surrounding area. Various sectors, industries or niches of the market are represented by each community. There is a wide variety of specializations amongst the communities, with almost every community having a different main type of contract. Of course, the biggest specializations are represented by the disconnected components, which have very limited or even zero competitive interactions with firms outside the community. Some communities have the same contract specialization (*C7* and *C11*; *C12* and *C17*) but differ on region activity. For example, *C7* and *C11* are both communities where the main subject of the contracts is construction, but *C11* is highly specialized in a single region (Madeira) thus the large distance between the communities in the graphical representation. Communities *C7* and *C10* are, by a considerable margin, the largest communities in the network. This could possibly indicate that *Construction* and *Medical & Pharmaceutical* are the two most competitive sectors in Portugal within the public procurement market.

Occasionally, firms can form densely connected sub-graphs (e.g., *C16*, *C17*, *C20*, *C22*). These shapes may serve as a first warning sign that a community poses a higher risk of engaging in fraudulent activities such as collusion and procurement manipulation. It could indicate that several firms in the same community interact frequently and closely, and even reach agreements to secure contracts and manipulate the process. Additionally, these structures could signal that a few firms dominate a certain market, making it easier to attempt engagement in illegal practices. That being the case, we need to explore additional metrics to categorize each community based on their behavior. This will help identify the groups of firms that might be worth investigating and looking further into.

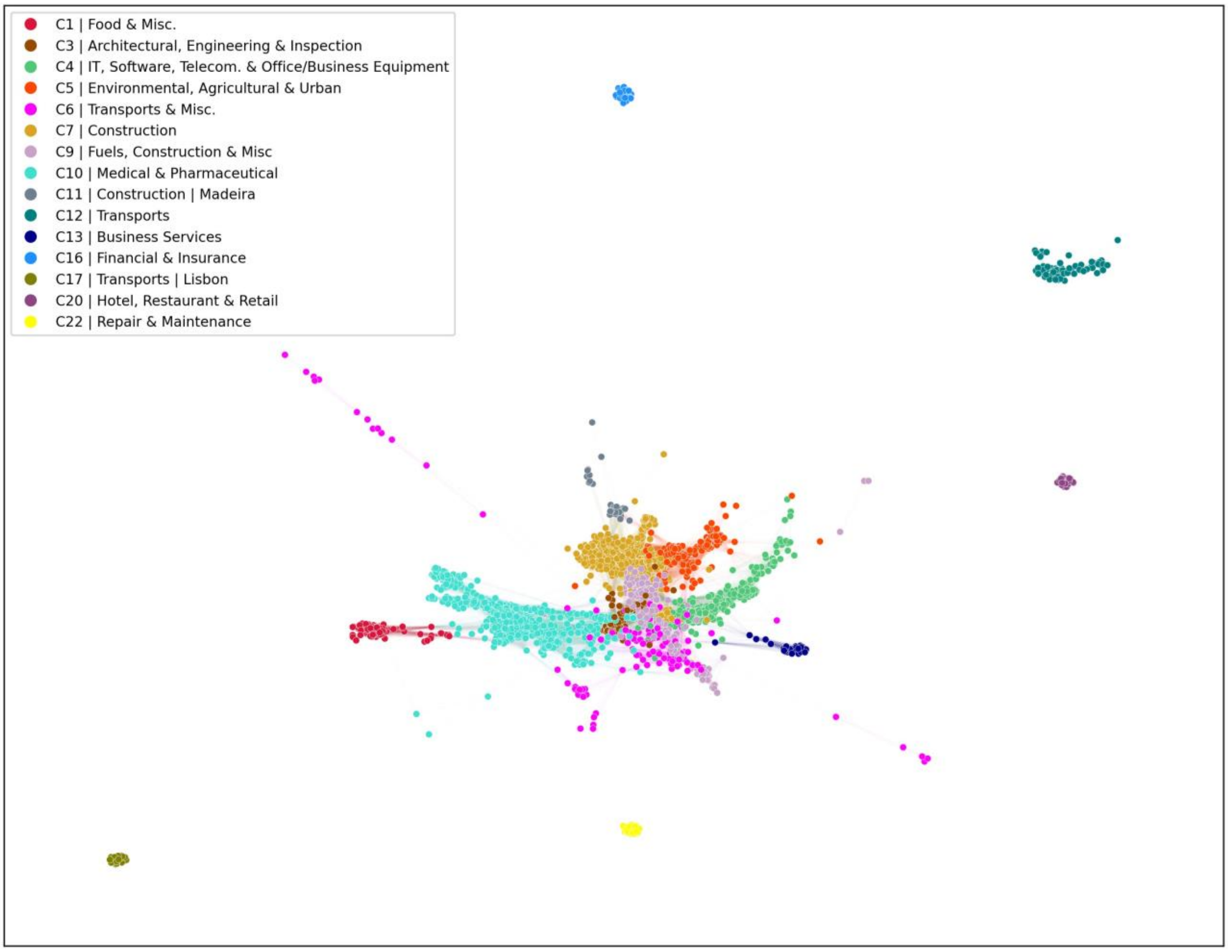


Figure 7 | **Firm-firm co-competition network**. Firms in the same community present a higher likelihood of cooperation and collaboration, since these will be firms which are more likely to bid together as a group, based on co-occurrence values. Represented are the 15 communities with the biggest size (number of firms). The network is composed of the core of active firms and edges with a statistically significant cosine similarity. This section of the network has a modularity of 0.74.

4.2. Activity Diversity

To further investigate and potentially find clues that could be associated with illegal behavior, we look at regional and contract diversity: regional referring to the Portuguese regions where each community performs their activities; and contract being the type of contract – defined by the CPV code – where communities bid on. Particularly, we'll be using the previously described Simpson's Index as the diversity Index.

While a firm with low regional and contract diversity could simply be explained by high specialization in both area of execution and industry sector, a group of firms which presents low diversity in both measurements could raise concerns. This lack of diversity within a community may be an indication of the circumstances in which businesses might coordinate and work together to achieve particular goals. It can imply collaboration or anti-competitive activity. For instance, these businesses may be collaborating to control a certain market or area, thus reducing competition, and possibly participating in dishonest or illegal tactics.

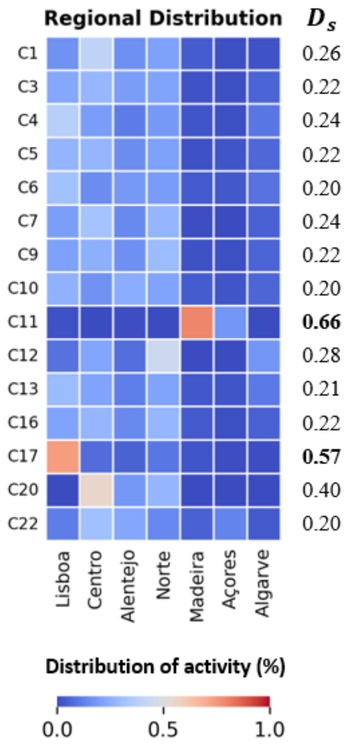
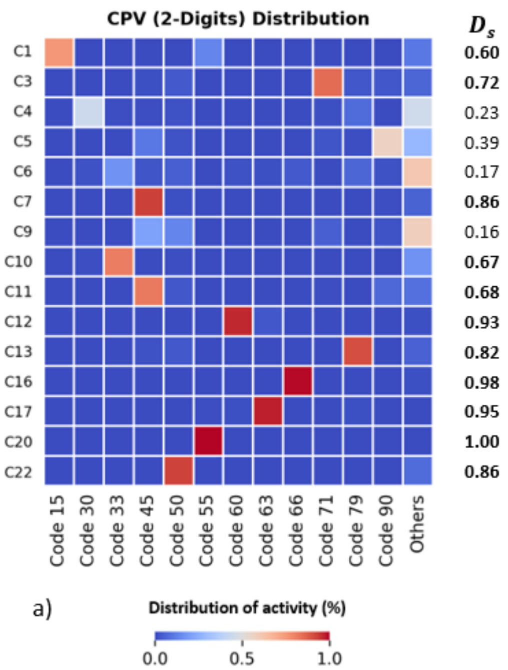


Figure 8 | **Regional distribution**. The figure displays the regional distribution of the bids made by the fifteen biggest communities in the pre-defined Portuguese regions. The Simpson's Index (D_s) of each community is presented on the right.

Figure [8](#) demonstrates the empirical regional distributions for the largest communities, while figure [9a](#) presents the contract type distributions. The red spectrum colors designate a high proportion of bids whereas the blue spectrum colors represent a low proportion of bids. Due to visualization purposes, figure [9a](#) does not present all the possible 2-digit CPV codes. The twelve codes available were selected because they were the ones that best represented the biggest communities. The remaining CPV codes are stacked into the "Others" category. It is worth mentioning that the Simpson's index calculation was made using every single one of the possible 45 different CPV codes. Figure [10](#) compares both dimensions in a single graph, with all 23 communities represented.

As mentioned in the previous section, a big portion of regional information was missing. The diversity index presented in figure [8](#) was computed utilizing only contracts whose region was not missing. This means that this figure is not totally representative of reality, but it is the best approximation given the data available. This affects community *C20* the most. A huge portion (99%) of its regional information was missing; thus, the Simpson's index value is probably not very realistic. Besides, communities were formed based on co-occurrence similarities, not region of activity. Hence, the fact that firms *C20* present so many missing values may be due to the nature of the contracts they specialize in.



Code	Subject
15	Food, beverages, tobacco, and related products
30	Office and computing machinery, equipment and supplies except furniture and software packages
33	Medical equipment, pharmaceuticals, and personal care products
45	Construction work
50	Repair and maintenance services
55	Hotel, restaurant, and retail trade services
60	Transport services (excl. Waste transport)
63	Supporting and auxiliary transport services; travel agencies services
66	Financial and insurance services
71	Architectural, construction, engineering, and inspection services
79	Business services: law, marketing, consulting, recruitment, printing, and security
90	Sewage, refuse, cleaning, and environmental services

b)

Figure 9 | **Contract type distribution**. Panel a) displays the CPV distribution of the bids made by the 15 biggest communities. The Simpson's Index (D_s) of each community is presented on the right of panel a). For visualization purposes, we use the CPV 2-digit code (e.g., Code 15) instead of division names. Each code-subject correspondence can be found in the table provided in panel b).

From these two metrics we can observe that out of the biggest communities, only two - *C11* and *C17* – present a red colored region, resulting in the two highest Simpson indexes. The same trend does not apply to contract type diversity, since 11 out of the 15 biggest communities present a high Simpson's index. This indicates that, in Portugal, firms are more likely to specialize in a particular industry sector rather than region of activity. In fact, we can find a big clustering of communities on the left upper quadrant in figure 10, representing firms with high levels of specialization and lower levels of regional agglomeration. On the other hand, the work of Lyra et. al. (2021) provides findings that indicate that most of the communities of firms in the state of Ceará (Brazil) present low levels of specialization and agglomeration. The high specialization in Portugal can be associated with a more mature market. Firms compete more fiercely in developed marketplaces. Market saturation may have caused a focus on specialization and distinction. These differences could also be explained by economic differences between Brazil and Portugal. For instance, Brazil's larger size and vast number of regions may entice businesses to diversify across industries and locations to take advantage of different opportunities. Communities in the top right quadrant have both high specialization in contract type and geographical activity. To the community degree, these two factors combined can be a useful resource to identify groups of firms which may be engaging in some type of fraudulent scheme (Lyra et. al. 2021).

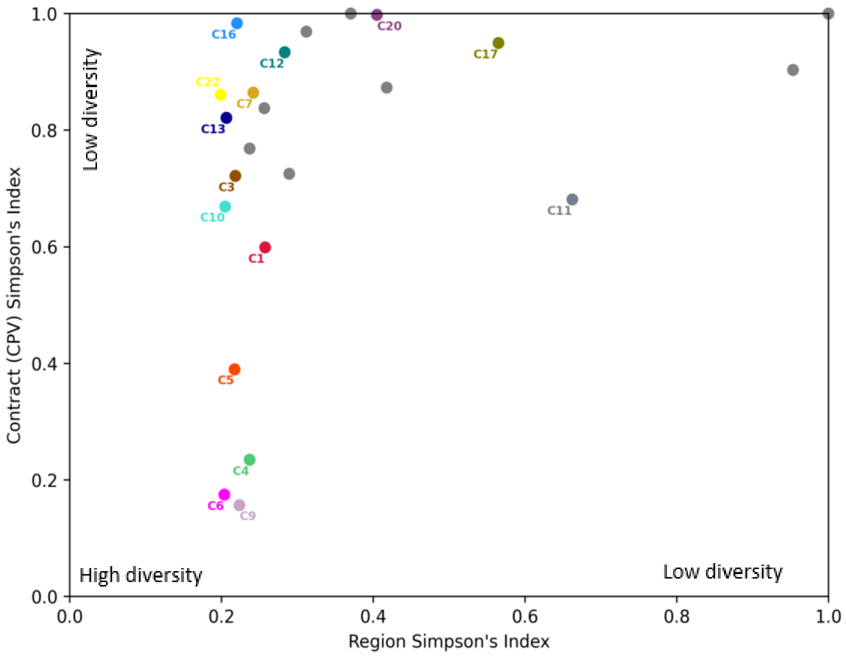


Figure 10 | **Comparison between contract type and regional diversity**. The colored communities represent the fifteen biggest communities, and the remaining are colored in gray.

This combination suggests that firms are controlling a specific market in a specific area. It could also suggest that transparency is limited in procurement processes as firms may find it simpler to persuade local decision makers.

It is crucial to emphasize that these indicators do not guarantee the existence of corruption or malpractice within the communities. However, these red flags can be useful in helping audit officials select firms to investigate. Instead of randomly sampling firms for investigation, officials may rely on these results to facilitate the selection process. Having that said, it is understood that to confirm any allegations of misconduct inside these communities, additional research is required.

Overall, these indicators helped figure out the level of specialization and consolidation of each community. Particularly, communities that belong to the top right quadrant stand out - *C11* and *C17* out of the biggest communities - and even more evidently *C8* (farthest top right) and *C23* out of the smallest communities. To a deeper extent, community *C8* only specializes in *Transport Services* contracts in the *Centro* (Center) region of Portugal, and community *C23* highly specializes in *Education and training services* contracts in the *Norte* (North) region. These four communities reveal low levels of diversity in both dimensions, which announces potential for collusion or other illegal methods amongst the firms in each group, thus possibly requiring further investigation by audit regulators.

4.3. SINGLE BIDDING

To explore more thoroughly the possibility of the communities engaging in market rigging practices, we explore each community's inclination to take part in “single bidder” contracts. “Single bidding” in the public sector refers to contracts where only a single firm has bid. It can often signal market manipulation, as it is associated with lack of competition and transparency. When only a single bid is made, it leads to questioning if the procedure has been rigged to keep out possible rivals, which could reduce fairness and openness. It can also lead to inflation of contract prices and consequently harm the public interest. Hence, we start by researching the average number of times each firm is a single bidder on a tender, per community. Figure 11 displays the findings for all communities found in the co-competition network. This graph can help identify groups of firms that attract attention for either one of two reasons: a high value of single bidding within a community, which may signify informational advantage over competitors or a sign of lack of competition in the market; a low value of single bidding, which can imply some sort of coordination - firms consistently bidding on the same tenders - to secure contracts.

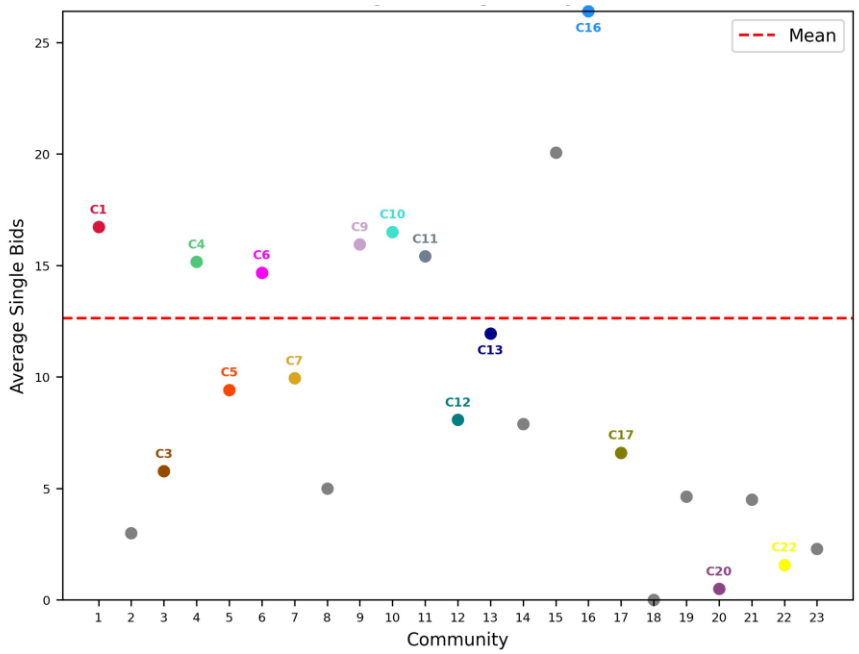


Figure 11 | **Average single bids per community**. Each value represents the average number of times in a community where a firm has been a single bidder (e.g., A value of 20 would mean that a firm from that community would have been a single bidder, on average, 20 times).

Overall, the average number of single bids per community obtained is 12.6. Out of the largest communities, *C16* presents a particularly high value of single bids, equaling a bit more than twice the average. On the other end, *C20* and *C22* have very low values of single bidding. Moreover, community *C18* presents the lowest value (zero) of all communities, meaning that there has not been an instance where a firm from this community has been a single bidder.

To strengthen our investigation, we also compute the average number of bidders per tender in each community. This allows us to identify coherences between the community size and tender size, which could help understand the existence of coordination between firms. Figure 12 displays these values, normalized. These were obtained by gathering the average number of firms in a tender per community and dividing them by community size – values can be found in Table C in the appendix. With this approach, it is to be expected that larger communities present values closer to zero while smaller communities present higher values. Nonetheless, a value close to one would mean that firms within a community usually bid on tenders the same size as the community itself. This information combined with a very low level of single bidding suggests a higher possibility that a group of firms might be coordinating their bids to get an advantadge.

The results show that communities *C8* and *C23* tend to bid in tenders with the same size as the community itself. Communities *C2* and *C21* tend to bid in tenders where the number of participants is well over the size of the community. These are the cases that present the most visible deviations, and should be considered for further investigation. Although the value of communities *C18* and *C20* are not much out of the norm, they are among the highest within

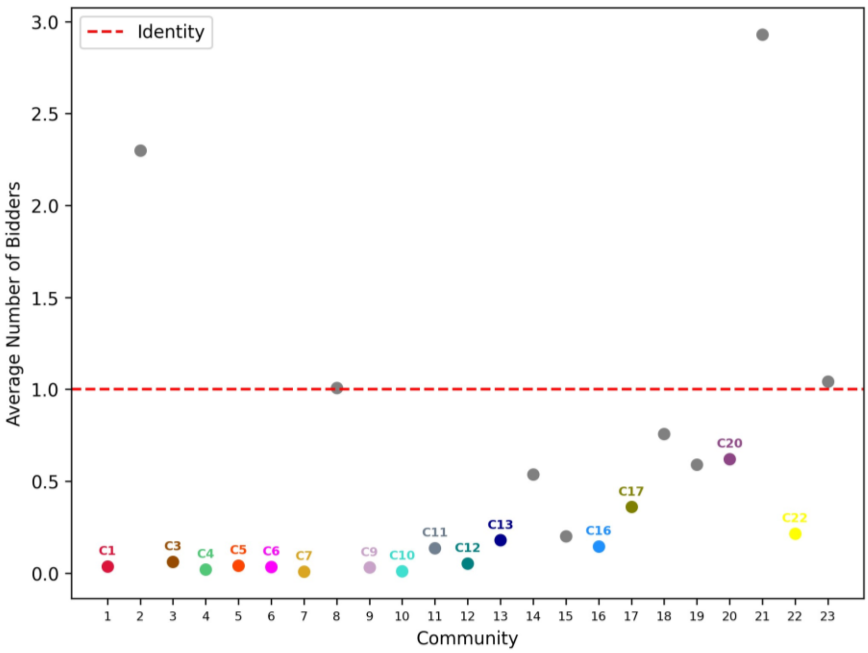


Figure 12 | **Average number of bids per tender per community**. This plot displays the average number of tender sizes in which firms within a community typically participate in. Each value was normalized by the community size. The threshold designating the community's size is depicted by the horizontal red line (i.e., a value above the red line would mean that firms in the community usually bid on tenders with more participants than the size of the community).

the range of values for these communities. Besides, there are also some other factors that should be taken into account. These two communities present the two lowest values of single bidding (figure 11) and the two highest values of average tender size (Table C for reference). Peculiarly, community *C20* exhibits an average tender size much larger than any other community (22.3 while the second highest is 9.8). A difference this big could have a deeper meaning. The blend of these metrics indicates that there is a chance that these groups of firms might be collaborating and engaging in fraudulent tactics to their benefit.

To reiterate, all the indicators that were discussed do not guarantee the presence of corruption in any of the communities. They merely serve as a tool to help select firms for auditing.

5. CONCLUSIONS AND FUTURE WORKS

The work presented in this thesis looked to explore and discover valuable information from a huge set of public procurement contracts issued by all the Portuguese regions. Specifically, data mining and network analysis techniques were used to find patterns in communities of firms associated with corruption and fraud. A co-competition network was obtained using the similarity co-occurrence patterns of the firms, which comprises 2801 firms and 130,471 edges.

We follow up by identifying communities within the network. A total of 23 communities were found, achieving a modularity of 0.74. We found that these communities form groups of firms that share mainly the same span of contract subjects (industry sectors) as well as, in some cases, region of activity, in their public procurement processes. Furthermore, we investigate other indicators that can help identify communities to gain influence or control in the procurement process. For that end, we utilized a diversity index to measure each community's contract type and regional variety, as well as the average number of single biddings and average number of bidders in each tender. Additionally, we found that the Portuguese public sector market seems to be more specialized in comparison to Brazil, which can be associated with a more mature and competitive market.

According to the findings, communities *C20* and *C23* seem to be the clusters of firms which gather the greatest number of suspicious features. *C20*: this community exclusively specializes in the hotel, restaurant, and retail industry. It also presents a very low average value of single bids while having the biggest, by far, average tender size. Additionally, the average tender size is not too far from the community size. As previously stated, the combination of these factors points to possible coordination and/or collusion among the firms in the group. As far as region of activity, the number of missing values does not allow a precise answer. This means that this community can either have a very narrow or broad scope of geographical activity. Although the missing values can be correlated with the nature of the industry, they can also be considered as another factor to justify a more thorough investigation by audit regulators; *C23*: this community presents low levels of diversity, highly specializing in education services in the northern area of Portugal. In addition, it presents one of lowest average values of single bids while having an average tender size equal to the number of firms in the community. All these flags combined could possibly justify a posterior investigation by audit regulators, although they do not prove the existence of corruption.

The exploration and results obtained in this thesis are based on the work of Lyra et. al. (2021). It provides new insights from a different country - Portugal - from a larger temporal range and can serve as a comparison base for both countries, therefore contributing to the literature on the subject. Regarding restrictions encountered, the abundance of missing values on regional information clearly limits the veracity of the regional diversity for some communities, mainly *C20* and *C23*. In addition, the lack of access to an historical dataset where past corruption cases are identified and labeled hinders the capacity to justify relationships between firms, their patterns and influence in the network. On top of that, the network obtained was quite complex. Due to its size and lack of computational power, randomizations proved to take quite long to process, which limited the use of other statistical and randomization methods, as well as simply using a larger number of randomizations for a more consistent model.

As an extension of this thesis, future works should involve exploring how, over longer times, community interactions and network structure change; future change prediction models can be created using this data; study the impact of government policy changes on the network; investigate resilience and robustness of the network to recognize weak points and create plans to strengthen them; utilize social network analysis methods to comprehend not only the economic activity but also the social dynamics and relationships within communities of interest. In addition, to perform similar analysis in other countries could be useful to compare behaviors and possibly find common factors that can help fight corruption in the public procurement sector, and ultimately save billions in public expenditure.

BIBLIOGRAPHICAL REFERENCES

Abdou, A., Basdevant, O., David-Barrett, E., & Fazekas, M. (2022). *Assessing Vulnerabilities to Corruption in Public Procurement and Their Price Impact* (No. 2022/094). International Monetary Fund.

Arrowsmith, S., Treumer, S., Fejø, J., & Jiang, L. (2011). *Public procurement regulation: an introduction*. University of Nottingham.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10), P10008.

Browne, C.M. (2017). Benchmarking Public Procurement. Foreword. World Bank. https://documents1.worldbank.org/curated/en/121001523554026106/Benchmarking-Public-Procurement-2017-Assessing-Public-Procurement-Regulatory-Systems-in-180-Economies.pdf

Carneiro, D., Veloso, P., Ventura, A., Palumbo, G., & Costa, J. (2020). Network analysis for fraud detection in portuguese public procurement. In *Intelligent Data Engineering and Automated Learning–IDEAL 2020: 21st International Conference, Guimaraes, Portugal, November 4–6, 2020, Proceedings, Part II 21* (pp. 390-401). Springer International Publishing.

Cheng, T., Liu, T., Meng, L., & Wang, C. (2017). The analysis of water project bid rigging behavior based on complex network. In *2017 International Conference on Applied Mathematics, Modeling and Simulation (AMMS 2017)* (pp. 406-410). Atlantis Press.

CIPFA. (2020). *Fraud and Corruption Tracker National Report*. The Chartered Institute of Public Finance and Accountancy. https://www.cipfa.org/services/cipfa-solutions/fraud-and-corruption/fraud-and-corruption-tracker

Comey, J. B. (2016). Oversight of the Federal Bureau of Investigation. Federal Bureau of Investigation. https://www.fbi.gov/news/testimony/oversight-of-the-federal-bureau-of-investigation

Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs. (n.d.) *Public procurement*. European Commission, Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs. https://single-market-economy.ec.europa.eu/single-market/public-procurement_en

European Commission. (2018). *Public Procurement Indicators 2018*. Retrieved from https://ec.europa.eu/docsroom/documents/48156 - accessed October 2023

European Commission. (2022). *Public procurement - Study on administrative capacity in the EU Portugal Country Profile.* https://ec.europa.eu/regional_policy/sources/policy/how/improving-investment/public-procurement/study/country_profile/pt.pdf

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

Fazekas, M., Cingolani, L., & Tóth, B. (2018). Innovations in objectively measuring corruption in public procurement. *Governance indicators. Approaches, progress, promise*, 154-185.

Fierăscu, S. I. (2017). *THE NETWORKED PHENOMENON OF STATE CAPTURE: Network Dynamics, Unintended Consequences, and Business-Political Relations in Hungary* [Doctoral dissertation, Central European University].

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, 486(3-5), 75-174.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12), 7821-7826.

Gottschalk, P., & Stanislas, P. (Eds.). (2017). *Public Corruption: Regional and National perspectives on procurement fraud*. CRC Press.

Grassi, R., Calderoni, F., Bianchi, M., & Torriero, A. (2019). Betweenness to assess leaders in criminal networks: New evidence using the dual projection approach. *Social Networks*, *56*, 23-32.

Gulati, R., Wohlgezogen, F., & Zhelyazkov, P. (2012). The two facets of collaboration: Cooperation and coordination in strategic alliances. *Academy of Management Annals*, 6(1), 531-583.

IMPIC. (2021). *Relatório anual de contratação pública em Portugal 2021*, 11-12. https://www.base.gov.pt/Base4/media/burjhsro/relat%C3%B3rio-anual-da-contrata%C3%A7%C3%A3o-p%C3%BAblica-2021.pdf

INE. (2015). *NUTS 2013 : as novas unidades territoriais para fins estatísticos*. Instituto Nacional de Estatística. Retrived from https://www.ine.pt/xurl/pub/230205992

International Monetary Fund. (2019). *Fiscal Monitor: Curbing Corruption*, 44-45. Retrieved from https://www.imf.org/en/Publications/FM/Issues/2019/03/18/fiscal-monitor-april-2019

LaCascia, H., & Kramer, M. (2023). Vulnerabilities of ICT Procurement to Fraud and Corruption. *Governance and the Digital Economy in Africa Technical Background Paper Series*. World Bank. http://hdl.handle.net/10986/40274

Lember, V., Kattel, R., & Kalvet, T. (Eds.). (2013). *Public procurement, innovation and policy: International perspectives*. Springer Science & Business Media.

Lynch, J., & Angel, J. (2013). Public procurement: Principles, categories and methods. *Procurement Classroom Series*, *2*.

Lyra, M. S., Curado, A., Damásio, B., Bação, F., & Pinheiro, F. L. (2021). Characterization of the firm–firm public procurement co-bidding network from the State of Ceará (Brazil) municipalities. *Applied Network Science*, *6*(1), 1-10.

Lyra, M. S., Pinheiro, F. L., & Bacao, F. (2022). Public Procurement Fraud Detection: A Review Using Network Analysis. In *Complex Networks & Their Applications X: Volume 1, Proceedings of the Tenth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2021 10* (pp. 116-129). Springer International Publishing.

Magakwe, J. (2022). The Root Causes of Corruption in Public Procurement: A Global Perspective. In *Corruption-New Insights*. IntechOpen.

Marshall, R. C., & Marx, L. M. (2014). *The economics of collusion: Cartels and bidding rings*. Mit Press. https://doi.org/10.7551/mitpress/9011.001.0001

Mendes, S. M., & Camões, P. J. (2005). Corruption and political institutions in Portugal: an exploratory analysis. https://hdl.handle.net/1822/3802

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23), 8577-8582

Newman, M. E. J. (2010). *Networks: An introduction*. Oxford University Press.

Newman, M. E., Cantwell, G. T., & Young, J. G. (2020). Improved mutual information measure for clustering, classification, and community detection. *Physical Review E*, *101*(4), 042304.

OECD. (2005). *Fighting Corruption and Promoting Integrity in Public Procurement*. OECD Publishing. https://doi.org/10.1787/9789264014008-en

OECD. (2010). *Public Procurement for Sustainable Development* [Brochure]. https://www.oecd.org/gov/ethics/Public-Procurement-for%20Sustainable-and-Inclusive-Growth_Brochure.pdf

OECD. (2014). *OECD Foreign Bribery Report: An Analysis of the Crime of Bribery of Foreign Public Officials*. OECD Publishing. https://doi.org/10.1787/9789264226616-en

OECD. (2016). *Preventing Corruption in Public Procurement* [Brochure]. https://www.oecd.org/gov/ethics/Corruption-Public-Procurement-Brochure.pdf

OECD. (2023). *Public procurement performance: A framework for measuring efficiency, compliance and strategic goals* (No. 36). OECD Public Governance Policy Papers. Paris: OECD Publishing. https://doi.org/10.1787/0dde73f4-en

OECD. (n.d.). *Integrity in Public Procurement*. https://www.oecd.org/gov/public-procurement/integrity/

Piccolo, S., Lehmann, S., & Maier, A. (2018). Design process robustness: A bipartite network analysis reveals the central importance of people. *Design Science*, 4, E1, 5-14.

Rose-Ackerman, S. (1999). *Corruption and Government: Causes, Consequences, and Reform*. Cambridge University Press. https://doi.org/10.1017/CBO9781139175098

Sidorov, G., Gelbukh, A., Gómez-Adorno, H., & Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18(3), 491-504.

Simpson, E. H. (1949). Measurement of diversity. *nature*, 163(4148), 688-688.

Siyal, S., & Xin, C. (2020). Public Procurement. *Global encyclopedia of public administration, public policy, and governance. Springer, Cham.*

Søreide, T. (2002). *Corruption in public procurement. Causes, consequences and cures*. Chr. Michelsen Intitute.

Stansbury, N. (2005). Exposing the foundations of corruption in construction. *Global corruption report*, *2005*, 36-55.

Thai, K. V. (2001). Public procurement re-examined. *Journal of public procurement*, *1*(1), 9-50.

Transparency International. (2021). *Global Corruption Barometer: EU 2021*. https://images.transparencycdn.org/images/TI_GCB_EU_2021_web_2021-06-14-151758.pdf

Wachs, J., & Kertész, J. (2019). A network approach to cartel detection in public auction markets. *Scientific reports*, *9*(1), 10818.

Wachs, J., Fazekas, M., & Kertész, J. (2021). Corruption risk in contracting markets: a network science perspective. *International Journal of Data Science and Analytics*, *12*, 45-60.

Walker, H., & Brammer, S. (2009). *Sustainable procurement in the United Kingdom public sector*. *Supply Chain Management: An International Journal, 14(2), 128–137.*

Wells, J. (2013). Corruption and collusion in construction: a view from the industry. *Corruption, grabbing and development: Real world challenges*, 23-34.

Williams-Elegbe, S. (2012). *Fighting corruption in public procurement: a comparative analysis of disqualification or debarment measures*. Bloomsbury Publishing.

World Bank. (2020, March). *Global Public Procurement Database: Share, Compare, Improve!* https://www.worldbank.org/en/news/feature/2020/03/23/global-public-procurement-database-share-compare-improve

Zhou, P., Hu, X., & Zhu, J. (2016). The normalization of occurrence and Co-occurrence matrices in bibliometrics using Cosine similarities and Ochiai coefficients. Journal of the *Association for Information Science and Technology*, 67(11), 2805-2814.

Zhu, J., Wang, B., Li, L., & Wang, J. (2020). Bidder network community division and collusion suspicion analysis in chinese construction projects. *Advances in Civil Engineering*, *2020*, 1-14.

APPENDIX

District	Region
Lisboa	Lisboa
Faro	Algarve
Açores	Açores
Madeira	Madeira
Viana do Castelo	Norte
Braga	Norte
Porto	Norte
Bragança	Norte
Vila Real	Norte
Aveiro	Centro
Viseu	Centro
Guarda	Centro
Coimbra	Centro
Leiria	Centro
Castelo Branco	Centro
Setúbal	Alentejo
Évora	Alentejo
Portalegre	Alentejo
Santarém	Alentejo
Beja	Alentejo

Table A - | **Portuguese district-region correspondence**

Community	Number of firms
C1	146
C2	2
C3	130
C4	286
C5	187
C6	135
C7	732
C8	2
C9	179
C10	663
C11	36
C12	83
C13	35
C14	9
C15	14
C16	27
C17	25
C18	13
C19	8
C20	36
C21	2
C22	44
C23	7

Table B | **Community Sizes.**

Community	Average tender size	Community Size	Normalized tender size
1	5.356187	146	0.036686
2	4.6	2	2.3
3	8.071491	130	0.062088
4	5.686368	286	0.019882
5	7.725072	187	0.041311
6	4.694264	135	0.034772
7	6.059783	732	0.008278
8	2.017857	2	1.008929
9	5.504621	179	0.030752
10	7.654086	663	0.011545
11	4.919251	36	0.136646
12	4.357791	83	0.052504
13	6.281311	35	0.179466
14	4.830688	9	0.536743
15	2.809406	14	0.200672
16	3.95599	27	0.146518
17	9	25	0.36
18	9.834783	13	0.756522
19	4.715953	8	0.589494
20	22.33913	36	0.620531
21	5.859375	2	2.929688
22	9.466667	44	0.215152
23	7.300885	7	1.042984

Table C | **Normalized tender sizes per community.**

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa