Universidade Federal da Paraíba
Centro de Tecnologia
Programa de Pós-Graduação em Engenharia Mecânica
Mestrado – Doutorado

IDENTIFICAÇÃO DE POSSÍVEIS REDES DE CONLUIO EM LICITAÇÕES PÚBLICAS UTILIZANDO TEORIA DOS GRAFOS, CLUSTERIZAÇÃO E PSO

por

Cecília de Freitas Vieira Couto

Dissertação de Mestrado apresentada à Universidade Federal da Paraíba para obtenção do grau de Mestre.

João Pessoa - Paraíba **Março, 2023**

CECÍLIA DE FREITAS VIEIRA COUTO

IDENTIFICAÇÃO DE POSSÍVEIS REDES DE CONLUIO EM LICITAÇÕES PÚBLICAS UTILIZANDO TEORIA DOS GRAFOS, CLUSTERIZAÇÃO E PSO

Dissertação de mestrado apresentada ao curso de Pós-Graduação em Engenharia Mecânica da Universidade Federal da Paraíba, em cumprimento às exigências para obtenção do Grau de Mestre.

Orientador: Prof. Dr. Lucidio dos Anjos Formiga Cabral

João Pessoa - Paraíba **Março, 2023**

Catalogação na publicação
Seção de Catalogação e Classificação

C871i Couto, Cecília de Freitas Vieira.
Identificação de possíveis redes de conluio em licitações públicas utilizando teoria dos grafos, clusterização e PSO / Cecília de Freitas Vieira Couto. - João Pessoa, 2023.
98 f. : il.

Orientação: Lucidio dos Anjos Formiga Cabral.
Dissertação (Mestrado) - UFPB/CT.

1. Licitações públicas. 2. Fraudes em licitações. 3. K-means. 4. Detecção de conluios. I. Cabral, Lucidio dos Anjos Formiga. II. Título.

UFPB/BC CDU 351.712(043)

Elaborado por GRACILENE BARBOSA FIGUEIREDO - CRB-15/794

IDENTIFICAÇÃO DE POSSÍVEIS REDES DE CONLUIO EM LICITAÇÕES PÚBLICAS UTILIZANDO TEORIA DOS GRAFOS, CLUSTERIZAÇÃO DE DADOS E META-HEURÍSTICAS DE OTIMIZAÇÃO

por

CECÍLIA DE FREITAS VIEIRA COUTO

Dissertação aprovada em 23 de março de 2023

Prof. Dr. LUCÍDIO DOS ANJOS FORMIGA CABRAL
Orientador – UFPB

Prof. Dr. SANDRO MARDEN TORRES
Examinador Interno – UFPB

Prof. Dr. GILBERTO FARIAS DE SOUSA FILHO
Examinador Externo – UFPB

Prof. Dr. MARÇAL ROSAS F LIMA FILHO
Examinador Externo – UFPB

João Pessoa - PB *2023*

AGRADECIMENTOS

Agradeço aos meus pais, Simone e Luiz Cláudio, que me incentivaram e me forneceram todo o apoio para que pudesse realizar este trabalho.

Agradeço aos professores do curso de Pós-Graduação em Engenharia Mecânica da Universidade Federal da Paraíba por todo conhecimento que me foi passado.

Agradeço pelas valiosas contribuições dadas pelo meu orientador, Lucidio Cabral, para a produção desta dissertação.

Agradeço, por fim, à banca examinadora da defesa do mestrado, Prof. Gilberto Farias de Sousa Filho, Prof. Sandro Marden Torres e Prof. Marçal Rosas Florentino Lima Filho, além do orientador Prof. Lucidio dos Anjos Formiga Cabral pela aceitação do convite.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

IDENTIFICAÇÃO DE POSSÍVEIS REDES DE CONLUIO EM LICITAÇÕES PÚBLICAS UTILIZANDO TEORIA DO GRAFOS, CLUSTERIZAÇÃO E PSO

RESUMO

As licitações públicas são um meio de contratação por meio do qual se busca garantir uma concorrência real entre os participantes, evitando-se, assim, a ocorrência de irregularidades. Apesar disso, não é incomum a identificação de inúmeros tipos de fraudes nesses processos licitatórios, dentre os quais se destaca a formação de conluio, que ocorre quando duas ou mais empresas se unem para fraudar uma licitação. Apesar de ser uma prática cometida a muitos anos, a detecção de conluios apresenta muitas dificuldades, especialmente devido à falta de ferramentas e de técnicas para auxiliar o processo de investigação. Ao longo dos anos, muitos métodos foram desenvolvidos visando auxiliar esse processo, mas devido a limitações técnicas, nenhum deles se estabeleceu em definitivo. Logo, diante disso, este trabalho teve por objetivo a elaboração de um método simples, mas com alto poder de utilização, para auxiliar a identificação de ocorrência de fraudes em licitações públicas. Foram utilizadas técnicas da teoria dos grafos e o algoritmo k-means otimizado pelo PSO para a identificação de relações suspeitas entre empresas. A metodologia proposta foi aplicada em dados de licitações públicas que ocorreram no estado da Paraíba entre os anos de 2014 e 2021, disponibilizados pelo Tribunal de Contas do Estado da Paraíba (TCE/PB) por meio do portal Sagres Online. Dessa forma, inicialmente o *dataset* obtido passou por um pré-processamento e por uma correção de incoerências. Em seguida, foram elaboradas visualizações gráficas utilizando grafos para exemplificar o uso da ferramenta. Após isso, os dados trabalhados foram agrupados e grupos de empresas com padrão de comportamento semelhantes foram identificados. Por fim, os resultados obtidos com a aplicação da teoria dos grafos foram unidos com os produtos da clusterização para a obtenção do resultado final desta dissertação, que mostrou o grande potencial do método proposto.

Palavras chaves – Licitações públicas, conluios, grafos, k-means, PSO.

IDENTIFICATION OF POSSIBLE NETWORKS OF COLLUSION IN PUBLIC PROCUREMENTS USING GRAPH THEORY, CLUSTERING AND PSO

ABSTRACT

Public tenders are a means of contracting which seeks to ensure real competition between participants, thus avoiding the occurrence of irregularities. Despite this, it is not uncommon to identify numerous types of fraud in these bidding processes, among which the formation of collusion stands out, which occurs when two or more companies come together to defraud a bid. Despite being a practice committed for many years, the detection of collusions presents many difficulties, especially due to the lack of tools and techniques to assist the investigation process. Over the years, many methods have been developed to help this process, but due to technical limitations, none of them has been definitively established. Therefore, in view of this, this work aimed to develop a simple method, but with high power of use, to help identify the occurrence of fraud in public tenders. Graph theory techniques and the k-means algorithm optimized by PSO were used to identify suspicious relationships between companies. The proposed methodology was applied to data from public tenders that took place in the state of Paraíba between the years 2014 and 2021, made available by the Court of Auditors of the State of Paraíba (TCE/PB) through the Sagres Online portal. Thus, initially the dataset obtained underwent pre-processing and a correction of inconsistencies. Then, graphical visualizations were elaborated using graphs to exemplify the use of the tool. After that, the worked data were grouped and groups of companies with similar behavior patterns were identified. Finally, the results obtained with the application of graph theory were united with the clustering products to obtain the final result of this dissertation, which showed the great potential of the proposed method.

Keywords – Public procurements, collusion, graphs, k-means, PSO.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

CAPÍTULO I

1. INTRODUÇÃO

A lei Nº 14.133 de 1º de abril de 2021 regulamenta atualmente as regras das licitações brasileiras. De acordo com tal lei, a utilização de licitações para as contratações busca garantir a seleção de propostas mais vantajosas à administração pública, certificar um tratamento isonômico entre os licitantes e uma competição justa e evitar contratações ou com sobrepreço, ou com preços extremamente baixos e inexequíveis.

Segundo a nova lei das licitações, os certames podem ser realizados em 5 modalidades diferentes, que são pregão, concorrência, concurso, leilão e diálogo competitivo (BRASIL, 2021). Cada uma dessas categorias apresenta diretrizes e características específicas.

Apesar de todas as regras e de todas as precauções estabelecidas tanto pela nova lei das licitações (lei 14.133 de 1º de abril de 2021) quanto pelo antigo regimento (lei 8.666 de 21 de junho de 1993), não é incomum a ocorrência de fraudes nos processos licitatórios.

Conforme dados do Portal da Transparência, foram realizadas no Brasil 16.701 licitações apenas no ano de 2022, que movimentaram aproximadamente R$ 392,85 bilhões de reais (PORTAL DA TRANSPARÊNCIA DO GOVERNO FEDERAL, 2022). Trata-se, portanto, de um vasto campo onde diversos tipos de crimes podem ser cometidos.

As fraudes que ocorrem durante um processo licitatório podem ser de diferentes naturezas, incluindo o direcionamento de licitações, a formação de conluios, o superfaturamento dos preços, das quantidades e/ou da qualidade, a realização do chamado jogo de planilhas etc.

Dentre essas práticas, destaca-se a formação de conluios, que ocorre quando duas ou mais empresas se unem para fraudar licitações públicas a partir, por exemplo, da divisão do mercado e/ou da definição de preços de propostas, de modo a provocar uma diminuição da concorrência real (O'SULLIVAN e SHEFFRIN, 2003). A ênfase nessa irregularidade

se deve à ausência de técnicas e de ferramentas que possam auxiliar sua detecção.

Para o combate a esse tipo de fraude, a metodologia aplicada atualmente pelos órgãos policiais brasileiros para identificar possíveis casos de irregularidade consiste em denúncias recebidas ou evidências circunstancias verificadas (CUIABANO et al, 2014). Assim, há uma grande impunidade no que se refere a esse tipo de crime (VELASCO et al, 2020).

Alguns métodos utilizando análise estatística, técnicas probabilísticas, mineração de dados, inteligência artificial etc. foram desenvolvidos ao longo dos anos para tentar fornecer uma base para uma detecção mais assertiva da formação de conluios em licitações, mas nenhuma das metodologias elaboradas se mostrou suficientemente boa a ponto de ter seu uso generalizado e disseminado (MODRUŠAN et al, 2021).

Diante desse cenário, fica evidente a necessidade do desenvolvimento de métodos mais eficazes para auxiliar a detecção da formação de conluios em licitações públicas. Para isso, este trabalho propõe uma nova metodologia que utiliza técnicas da teoria de grafos e algoritmos de clusterização para a elaboração de um método simples, mas com alto poder de utilização nos processos investigativos.

1.1. OBJETIVO GERAL

O objetivo geral desta dissertação foi desenvolver metodologias para o tratamento de dados de licitações públicas relacionadas à construção civil visando auxiliar o processo investigativo de possíveis irregularidades.

1.2. OBJETIVOS ESPECÍFICOS

Esta dissertação teve como objetivos específicos:

a) Selecionar e tratar um *dataset* com dados de licitações públicas;
b) Analisar e revisar os métodos utilizados até então para o estudo de irregularidades em licitações;
c) Elaborar uma metodologia de análise usando grafos;
d) Elaborar uma metodologia de análise usando clusterização;
e) Exemplificar como tais metodologias podem ser usadas na prática.

1.3. JUSTIFICATIVA

As formações de conluio, além de prejudicarem a livre concorrência, ainda provocam um aumento nos gastos das verbas públicas, que poderiam ser utilizadas para outro fim. Isso foi demonstrado por Smuda (2013), que verificou que a presença de conluio em uma licitação pode aumentar o valor dos itens contratados em até 20%.

Devido à falta de dados oficiais e unificados, não é possível determinar quanto é perdido por ano no Brasil devido às formações de conluio. Acredita-se, porém, que se trata de uma alta quantia, que poderia estar sendo investida na educação, na saúde, na segurança pública etc. Em um país como o Brasil, que ocupa a 87º posição entre 191 países no Índice de Desenvolvimento Humano (IDH), na lista gerada pela ONU para os anos de 2021 e 2022, com o índice de 0.754 (PNUD, 2022), os valores desviados apresentam um elevado custo social.

A formação de conluios em licitações é considerada crime pela Lei Nº 14.133/2021. Segundo o Art. 337-F dessa legislação, trata-se de uma infração "Frustrar ou fraudar, com o intuito de obter para si ou para outrem vantagem decorrente da adjudicação do objeto da licitação, o caráter competitivo do processo licitatório", o que prevê pena de reclusão de 4 a 8 anos de prisão, além do pagamento de multa (BRASIL, 2021).

Apesar da definição legal, formações de conluio em licitações públicas são crimes difíceis de serem identificados. Isso ocorre devido à falta de ferramentas que possam ser utilizadas para auxiliar as investigações e à ausência de métodos comprovadamente eficazes para uma correta e precisa verificação da ocorrência do crime.

Ao longo dos últimos anos, muitos métodos foram propostos visando satisfazer tal necessidade. Porém, geralmente tais técnicas ou não apresentam resultados bons ao serem generalizados, ou necessitam de muitas informações ou de dados muito específicos para que possam ser calibrados, ou exigem um grande tempo para a identificação dos casos suspeitos, ou apresentam outros empecilhos para seu uso prático.

Assim sendo, torna-se clara a necessidade de desenvolvimento de métodos para otimizar e padronizar o processo de investigação de licitações públicas afim de identificar a formação de conluios entre empresas. Neste trabalho, foram utilizadas técnicas da teoria dos grafos e de clusterização para o desenvolvimento de uma nova e simplificada metodologia visando tal fim.

Escolheu-se trabalhar apenas com licitações relacionadas à construção civil, devido ao grande montante envolvido em tais operações. Além disso, ao filtrar os tipos de certames analisados, garante-se que as licitações estudadas apresentarão padrões de comportamento mais semelhantes, o que tende a produzir resultados melhores, especialmente no processo de clusterização.

Ademais, os esquemas de conluio são realizados entre empresas que atuam em um mesmo setor. Por isso, faz sentido limitar os certames para apenas uma área de aplicação. Apesar disso, a metodologia desenvolvida neste trabalho pode ser tranquilamente utilizada para análises de licitações de outras áreas.

1.4. METODOLOGIA

A metodologia desta dissertação busca descrever as etapas realizadas para a elaboração do modelo apresentado, sendo elas:

- Etapa 1: Estudo e análise dos métodos utilizados por outros autores;
- Etapa 2: Seleção e tratamento do *dataset*;
- Etapa 3: Desenvolvimento do método usando visualização de grafos;
- Etapa 4: Desenvolvimento do método usando clusterização.

Na primeira etapa, buscou-se averiguar quais metodologias já foram ou que estão sendo empregadas por outros autores visando a solução do problema proposto nesta dissertação. Com isso, objetivou-se identificar os métodos que já foram utilizados para que novas metodologias pudessem ser sugeridas.

Na etapa 2, foi selecionado o *dataset* que será utilizado no decorrer desta dissertação. O conjunto de dados que será trabalhado é composto por dados de licitações públicas que ocorreram no Estado da Paraíba entre 2014 e 2021. Essas informações foram disponibilizadas pelo Tribunal de Contas do Estado da Paraíba (TCE/PB) por meio do portal Sagres Online.

Ainda na segunda etapa, o *dataset* selecionado foi tratado com a aplicação de filtros, com a escolha de variáveis, com a geração de novos parâmetros e com a correção de alguns dados.

Também na fase 2, foram obtidas informações relativas às empresas que foram alvo de processos investigativos relacionados a irregularidades em licitações públicas pelo TCE/PB.

A terceira etapa, que foi a de aplicação do método que utiliza grafos, consistiu na geração de subgrafos que permitiam a visualização de informações do *dataset* estudado. Foi criado uma série de cenários explorando as diversas possibilidades de análises que podem ser realizadas.

Nos grafos gerados, foram aplicados conceitos da teoria dos grafos, como os de cliques máximas e de densidade de um grafo, para a retirada de informações e análise dos resultados alcançados.

Na quarta etapa, onde foi utilizado o método de clusterização que une o k-means com o PSO (*Particle Swarm Optimization*), foram inicialmente criados cenários de agrupamentos realizados nos dados analisados. Com essas informações e realizando uma análise dos dados, as empresas analisadas foram divididas em grupos que apresentam características semelhantes.

Por fim, os resultados alcançados com a clusterização usando o k-means em conjunto com o PSO foram relacionados com os obtidos pelo método que usa grafos para a obtenção do resultado final desta dissertação.

Para a realização deste trabalho, foram utilizados apenas dados públicos fornecidos pelo TCE/PB e que podem ser acessados através da página da internet do tribunal[1].

Os códigos utilizados nesta dissertação foram elaborados na linguagem de programação *python* e foram executados utilizando a ferramenta Google Colab. Para isso, foi utilizado um notebook da marca *Acer*, com 8 GB de memória RAM, um processador Intel Core i5 e uma placa de vídeo NVIDIA Geforce 920M.

[1] Disponível em: https://tce.pb.gov.br/.

CAPÍTULO II

2. REVISÃO BIBLIOGRÁFICA

A busca por métodos que possam auxiliar o tratamento e a investigação de fraudes em licitações não é recente. Ainda em 1956, Friedman estudou técnicas probabilísticas que pudessem determinar, com certa precisão, a ocorrência ou não de fraudes em certames a partir de dados como o valor estimado do contrato e o valor gasto de fato na contratação (FRIEDMAN, 1956).

Com o aumento da divulgação de dados públicos de licitações e com o avanço das tecnologias, os estudos acerca da formação de conluios e outras irregularidades em certames ganharam destaque e passaram a ser analisados sobre outros aspectos e com técnicas novas.

Signor et al (2020) utilizou métodos probabilísticos para propor um modelo capaz de determinar a ocorrência de um conluio em tempo real a partir de informações das licitações, como o número de concorrentes e o desconto oferecido por cada competidor. Dessa forma, seria possível verificar a ocorrência do ato criminoso antes de sua execução, poupando tempo e dinheiro da administração pública.

Para a elaboração de tal método probabilístico, Signor et al (2020) usou dados do Sistema de Fiscalização Integrada de Gestão (e-Sfinge), que é um conjunto de aplicativos mantidos pelo Tribunal de Contas do Estado de Santa Catarina (TCE/SC) e que contêm informações relativas à administração pública de tal local (TCE/SC).

O *dataset* que será utilizado neste trabalho também foi utilizado por Fraga (2017) em sua dissertação, que teve como objetivo a identificação de grupos de empresas suspeitas de realizar práticas anticompetitivas, tais como a formação de cartéis, em licitações ocorridas nas cidades do estado da Paraíba entre 2005 e 2016.

Fraga (2017) utilizou técnicas de mineração de dados, o algoritmo Apriori e regras de associação nos dados das licitações para identificar padrões suspeitos em ambientes de maior concorrência, padrões suspeitos em ambientes de menor concorrência, evidências de simulação de concorrência e/ou evidências de concentração regional.

Lima (2010) estudou a relação entre os descontos oferecidos pelos competidores de uma licitação e a quantidade de concorrentes do certame. Para isso, ele utilizou informações dos "Relatórios Finais" das licitações disponibilizados pelo Departamento Nacional de Infraestrutura de Transportes (DNIT).

Em seu trabalho, Lima (2010) concluiu que as licitações que tiveram mais de oito concorrentes apresentaram um desconto médio maior do que os certames com menos participantes. Dessa forma, nas licitações com mais de oito competidores é assumida a existência de uma competitividade real, o que implica em uma baixa probabilidade de ocorrência de irregularidades.

Em um estudo realizado mais recentemente, Lima (2021) desenvolveu um modelo para detecção de indícios de ocorrência de fraudes e de formação de conluios em licitações de obras públicas brasileiras a partir da classificação de publicações do Diário Oficial da União.

Para isso, o autor usou as técnicas de Processamento de Linguagem Natural (NLP) TF-IDF e Doc2Vec, além de modelos lineares clássicos, redes neurais profundas, *bottleneck* e BI-LSTM (LIMA, 2021).

Rodríguez et al (2022) testou a acurácia de onze algoritmos de *machine learning* na detecção de conluios em dados de licitações. Para isso, foram utilizados *datasets* do Brasil, da Itália, do Japão, da Suíça e dos Estados Unidos.

A partir dos testes realizados, Rodríguez et al (2022) verificou que apesar de pouco utilizado atualmente, o *machine learning* apresenta grande potencial de utilização para auxiliar investigações de irregularidades em licitações públicas.

Foremny e Anysz (2018) utilizaram análise de dados para investigar a formação de cartéis em 98 licitações ocorridas na Polônia com o objetivo de construir ou reconstruir estradas do país. Os autores fizeram uso de quatro indicadores, que foram o número de propostas das licitações, o valor das propostas dos certames, a quantidade de concorrentes das licitações e a quantidade de vezes que uma empresa ganhava uma licitação em uma determinada região geográfica.

Velasco et al (2020) apresentou o sistema de apoio à decisão (SAD), uma ferramenta utilizada na análise sistemática de compras públicas que permite que órgãos de segurança determinem prioridades na investigação a certas empresas e/ou pessoas físicas. O SAD utiliza algoritmos de mineração de dados, de pesquisa operacional, da teoria de grafos, de clusterização e de análise de regressão.

Em seu trabalho, Velasco et al (2020) ainda estabeleceu alguns padrões de risco acerca do comportamento de empresas que podem indicar a ocorrência de fraudes em processos licitatórios. Os padrões de riscos determinados pelos autores são:

- Licitações idênticas: ocorre quando duas empresas participam de licitações diferentes, sempre com os mesmos valores de proposta;
- Licitações combinadas: duas empresas participam de diferentes processos licitatórios juntas com valores propostos proporcionais ou diferentes de um valor fixo;
- Top perdedores: Uma determinada empresa participa de várias licitações, mas quase nunca (ou nunca) ganha, de modo a aparentar uma falsa concorrência;
- Parceiros comuns: Duas empresas que possuem um parceiro comum participam dos mesmos processos licitatórios;
- Endereços comuns: Empresas que declaram o mesmo endereço participam das mesmas licitações;
- Grupo econômico comum: Duas ou mais empresas que pertencem ao mesmo grupo econômico (ao mesmo dono) participam dos mesmos processos licitatórios;
- Dados cadastrais comuns: Empresas que participam de um processo licitatório online e apresentam o mesmo endereço IP, indicando a existência de alguma relação entre os proponentes.

Modrušan et al (2021) realizou uma revisão da bibliografia relacionada ao desenvolvimento de técnicas para identificação de formação de conluios em licitações públicas. Ao todo, os autores analisaram 23 trabalhos e verificaram uma tendência ao desenvolvimento de metodologias voltadas para a detecção precoce da fraude, se possível ainda durante o processo licitatório, ao invés da identificação posteriori do crime, durante uma investigação criminal.

Modrušan et al (2021) constatou ainda que a qualidade dos modelos desenvolvidos para a detecção de irregularidades nos processos licitatórios depende diretamente da qualidade do conjunto de dados utilizado e dos indicadores de corrupção adotados.

A partir das análises realizadas, os autores identificaram os tipos de *datasets* mais utilizados pelos pesquisadores, os resultados comumente buscados, as métricas de avaliação mais utilizadas etc. Modrušan al (2021) constatou que os métodos mais empregados nos trabalhos investigados foram regressões lineares e logísticas, redes neurais e algoritmos de Naive Bayes. Essas técnicas foram usadas tanto para a classificação dos dados quanto para a sua clusterização.

Para uma melhor análise e comparação dos trabalhos apresentados, na Tabela 2.1 é exibido um breve resumo acerca desses materiais.

Tabela 2.1 – Resumo dos trabalhos apresentados na revisão bibliográfica.

Autor	Ano	Método empregado	Objetivo
Friedman	1956	Métodos probabilísticos	Determinar a ocorrência ou não de fraudes em certames
Signor et al	2020	Métodos probabilísticos	Criar um modelo para identificar a ocorrência de conluio em tempo real
Fraga	2017	Mineração de dados, algoritmo Apriori e regras de associação	Identificar grupos de empresas suspeitas de realizar práticas anticompetitivas
Lima	2010	Análise de dados	Comparar o número de concorrentes com o valor do desconto de uma licitação
Lima	2021	Processamento de Linguagem Natural (NLP), modelos lineares clássicos, redes neurais profundas, *bottleneck* e BI-LSTM	Desenvolver um modelo para detecção de indícios de ocorrência de fraudes e de formação de conluios em licitações
Rodríguez et al	2022	*Machine Learning*	Detectar conluios em dados de licitações
Foremny e Anysz	2018	Análise de dados	Investigar a formação de cartéis em 98 licitações da Polônia.
Velasco et al	2020	Mineração de dados, pesquisa operacional, teoria dos grafos, clusterização de dados e análise de regressão	Apresentar o sistema de apoio à decisão (SAD)
Modrušan et al	2021	Revisão bibliográfica	Analisar os trabalhos produzidos para identificar a formação de conluios em licitações

Como pôde ser constatado nessa breve revisão bibliográfica realizada, a investigação de irregularidades em licitações vem inspirando pesquisas de diversas naturezas. Ao longo dos anos, foram aplicadas várias metodologias visando desenvolver métodos eficazes para solucionar tal problema. Desde métodos probabilísticos até técnicas de inteligência artificial, muitos são os procedimentos empregados na detecção de fraudes em licitações.

Neste trabalho, uma nova metodologia será sugerida para facilitar o processo investigativo e para melhorar a identificação de conluios em licitações. O método proposto

busca identificar relações suspeitas entre empresas a partir da aplicação das técnicas da teoria dos grafos e da clusterização de dados.

Utilizando a técnica elaborada neste trabalho, será possível identificar licitações com suspeita de ocorrência de irregularidades antes da contratação de empresas, minimizando o dano ao erário. Além disso, será possível também verificar empresas suspeitas de cometerem fraudes em certames e grupos de companhias que mantém relações duvidosas entre si.

O método proposto neste trabalho utilizará dados de licitações públicas, mas ele pode ser adaptado para outros conjuntos de informações. As variáveis analisadas também podem ser modificadas e adequadas para as situações analisadas. Os detalhes da metodologia desenvolvida podem ser visualizados nas próximas seções.

CAPÍTULO III

3. REFERENCIAL TEÓRICO

Como apresentado na metodologia, nesta dissertação foi desenvolvido um método para auxiliar o processo investigativo de licitações baseado na teoria dos grafos e na clusterização de dados. Para um melhor entendimento dos resultados que serão apresentados, faz-se necessário introduzir alguns conceitos e algumas informações. Sendo assim, nas seções a seguir serão realizadas breves explanações para uma melhor compreensão do restante deste trabalho.

3.1. GRAFOS

A seguir serão apresentadas algumas definições relacionadas à teoria dos grafos que serão necessárias para um melhor entendimento desta dissertação.

- Grafos: Matematicamente, um grafo simples pode ser descrito como uma estrutura discreta formada por um conjunto não vazio de vértices V e por um conjunto de arestas A, onde cada aresta de A é formada por um par de vértices distintos do conjunto V (PRESTES, 2016).

  Trata-se de uma estrutura de abstração muito útil para a representação e solução de problemas computacionais, por meio dos quais é possível representar relações de interdependência entre elementos de conjuntos (CARVALHO, 2020). Os grafos podem ser utilizados em diversas áreas de conhecimento, como a física, a química, a psicologia, as engenharias etc. (PRESTES, 2016).

  Na Figura 3.1.1 a seguir é exibido um exemplo de um grafo.

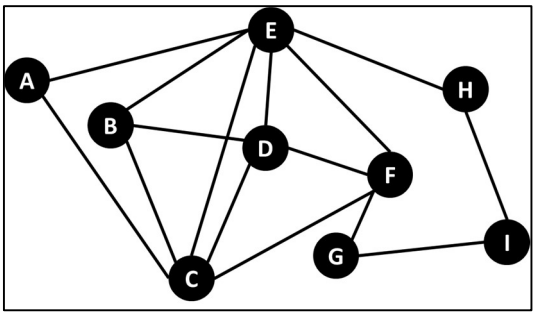


Figura 3.1.1 – Exemplo de um grafo.

- Subgrafos: Um subgrafo é uma parte de um grafo maior, o que significa que os conjuntos de vértices e de arestas do subgrafo são subconjuntos dos conjuntos de vértices e de arestas de um grafo maior (NOGUEIRA JÚNIOR, 2017). Na Figura 3.1.2 a seguir é possível ver um exemplo de um subgrafo da estrutura exibida na Figura 3.1.1.

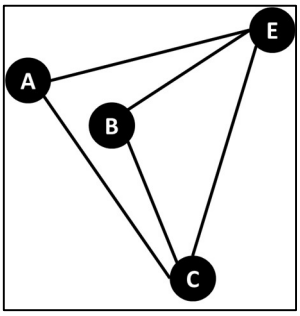


Figura 3.1.2 – Exemplo de um subgrafo do grafo da Figura 3.1.1.

- Grafos completos: Um grafo é dito completo se para cada par de vértices distintos, há uma aresta ligando-os (CARVALHO, 2020).
- Cliques: Uma clique de um grafo é um subgrafo completo dele. Isto é, dado um grafo, serão cliques todos os subgrafos completos dele (PRESTES, 2016).
- Cliques máximas: Uma clique de um grafo é dita máxima quando ela não é subgrafo de nenhum outro subgrafo completo (PRESTES, 2016). Isso significa que a clique máxima é, dentre as cliques de um grafo, a que apresenta o maior número de vértices (e consequentemente de arestas). Apesar de ser um conceito relativamente simples, a identificação de cliques máximas em grafos é um problema NP-difícil, o que significa que não existe um algoritmo conhecido que forneça uma solução ótima para o problema em tempo polinomial (ZÜGE, 2017).

  Na Figura 3.1.3 é exibida como exemplo a clique máxima do grafo da Figura 3.1.1.

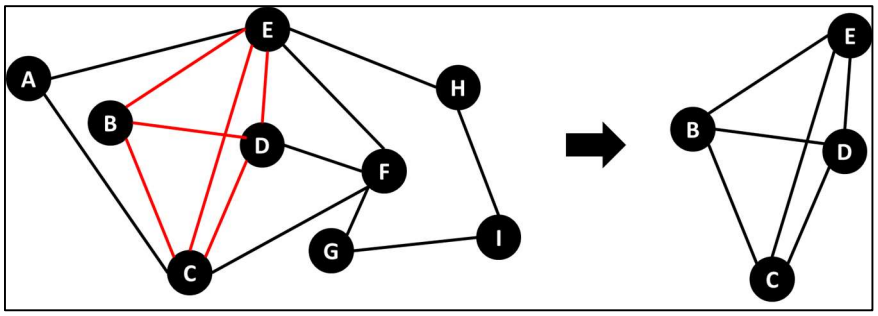


Figura 3.1.3 – Clique máxima do grafo da Figura 3.1.1.

- Densidade de um grafo: A densidade de um grafo é dada pela relação entre a quantidade de arestas que ele apresenta e a quantidade de arestas que um grafo completo com a mesma quantidade de vértices possui (DOS SANTOS e JUSTEL, 2015). Por exemplo, como o grafo mostrado na Figura 3.1.1 possui 15 arestas, mas deveria possuir 36 para ser um grafo completo, sua densidade vale aproximadamente 0,42.
- Clique quase máxima: As cliques quase máximas de um grafo são geradas a partir da expansão da sua clique máxima pela adição de nós que apresentam alto grau de relação com os nós da clique máxima, mas que por não estarem conectados com todos os elementos dessa, acabam não fazem parte dela. As cliques quase máximas não possuem limite de tamanho, mas são caracterizadas por possuírem altas densidades, com valores próximos a 1. A análise das cliques quase máximas permite identificar grupos de nós que possuem elevado grau de relação entre si.

  Na Figura 3.1.4 a seguir é apresentado um exemplo de uma clique quase máxima para o grafo da Figura 3.1.1. Esse subgrafo apresenta densidade de 0,90.

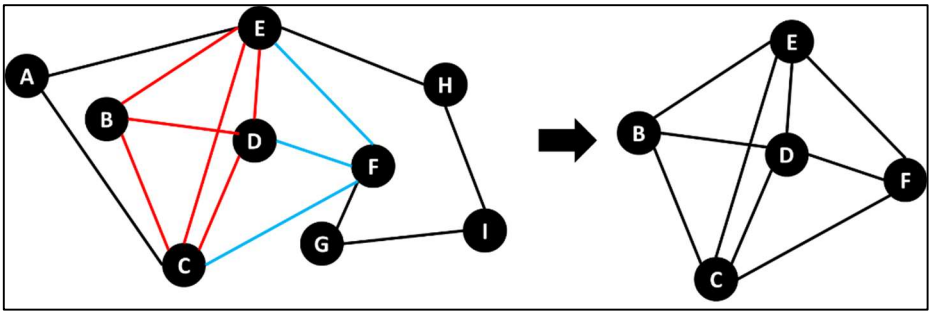


Figura 3.1.4 – Clique quase máxima do grafo da Figura 3.1.1.

- Grafos ponderados: São grafos que possuem pesos em suas arestas. Essas medidas podem representar custos, distâncias, tempos etc. (NOGUEIRA JÚNIOR, 2017).

3.2. CLUSTERIZAÇÃO USANDO K-MEANS E PSO

A clusterização é um importante método popularmente utilizado em diversas áreas. Ela pode ser entendida como uma técnica não supervisionada que visa separar um determinado conjunto de dados em diversos grupos com base em semelhanças existentes nas informações (PATEL et al, 2017).

Na computação, existem inúmeros algoritmos de clusterização de dados, sendo o k-means um dos mais conhecidos. Essa popularidade se justifica devido à facilidade de implementação e uso do algoritmo, à rapidez na sua execução e aos bons resultados apresentados por tal técnica (PATEL et al, 2017).

Para realizar a divisão dos dados nos grupos, o k-means usa um elemento para representar cada cluster, o centroide. Esses elementos são estabelecidos inicialmente de forma aleatória e são ajustados em seguida. Para realizar essa correção, a cada iteração são verificados os elementos mais similares (mais próximos) a cada um dos centroides e são formados os clusters (COELHO FILHO et al, 2013).

Em seguida, o centroide, que é representado pelo valor médio de todos os atributos do grupo que ele representa, é recalculado e uma nova organização nos dados é realizada. Esse processo ocorre, geralmente, durante um número fixo de iterações e/ou até que as posições dos centroides não mudem mais do que uma determinada tolerância (COELHO FILHO et al, 2013).

Apesar de ser amplamente utilizado, o k-means apresenta algumas desvantagens. Dentre essas, destaca-se a possibilidade de o algoritmo convergir para mínimos (ou máximos) locais ao invés de mínimos (ou máximos) globais, de modo a gerar um resultado que não é o melhor possível (PATEL et al, 2017).

Além disso, o resultado final do k-means depende do centroide selecionado inicialmente, o que ocorre de forma aleatória. Sendo assim, diferentes execuções do mesmo algoritmo em um mesmo conjunto de dados podem gerar resultados diferentes. Ademais, os grupos gerados pelo k-means são sensíveis a valores extremos (*outliers*), o que também pode comprometer o resultado final (PATEL et al, 2017).

Existem inúmeras técnicas que buscam contornar as dificuldades do k-means e, com isso, produzir melhores resultados. Dentre essas, destaca-se a utilização de meta-heurísticas que otimizam a geração dos centroides do k-means, melhorando a divisão dos dados.

Meta-heurísticas são estratégias de busca que exploram o espaço das soluções viáveis de um problema para evitar que ocorra o confinamento em mínimos ou máximos locais (BECCENERI, 2008).

Dentre as meta-heurísticas que podem ser aplicadas em conjunto com o k-means para melhorar a clusterização dos dados, tem-se a Otimização por Enxame de Partículas ou *Particle Swarm Optimization* (PSO) (COELHO FILHO et al, 2013).

Elaborado por Kennedy e Eberhart (1995), o PSO é um algoritmo de otimização que busca simular graficamente o comportamento de um bando de pássaros. Para isso, são utilizadas uma série de partículas que, após serem inicializadas de forma aleatória, se movimentam alterando sua velocidade para encontrar a melhor posição no espaço analisado.

Durante a execução do algoritmo, são realizadas iterações para atualizar as velocidades e as posições das partículas até que um ponto de parada seja alcançado (DORIGO et al, 2008). Para isso, a cada repetição, a velocidade de cada uma das partículas é modificada segundo a Equação 1 exposta a seguir.

$$v^{t+1} = wv^t + c_1 U_1^t (pb^t - x^t) + c_2 U_2^t (gb^t - x^t) \qquad \text{(Equação 1)}$$

Onde:

- v^{t+1} é a velocidade que a partícula irá assumir na iteração t + 1;
- w é o peso de inércia;
- v^t é a velocidade da partícula na iteração t;
- c_1 e c_2 são coeficientes de aceleração;
- U_1^t e U_2^t são números aleatórios uniformemente distribuídos no intervalo [0,1);
- pb^t é o valor do *personal best* da partícula;
- gb^t é o valor do *global best* do enxame;
- x^t é a posição que a partícula ocupa da iteração t.

A primeira parcela da soma da Equação 1, que também é chamada de inércia ou momento, funciona como uma memória da direção do voo anterior, de modo a impedir que a partícula sofra mudanças bruscas na sua velocidade (DORIGO et al, 2008).

O segundo termo da soma da Equação 1, conhecido como componente cognitivo, modela a tendência das partículas de voltar às melhores posições anteriormente encontradas. Por fim, o último elemento da soma da Equação 1, o componente social, adiciona ao desempenho de uma partícula o comportamento de suas vizinhas, indicando qual o padrão do grupo que deve ser obtido (DORIGO et al, 2008).

Na Equação 1, o peso de inércia *w* funciona como uma espécie de freio da velocidade. Como a cada iteração as partículas tendem a se aproximar do ponto de mínimo (ou máximo) global, elas são "freadas" para evitar que ultrapassem o local desejado, facilitando a convergência final do modelo.

Como se almeja que a partícula avance de forma rápida no início da execução do algoritmo e de modo mais devagar no final, o freio aplicado na velocidade não é, geralmente, uniforme. Para que tal situação ocorra, inicialmente é aplicado um valor maior para *w*, que vai diminuindo ao longo das iterações. Nesta dissertação adotou-se uma diminuição linear no valor de *w*.

Os coeficientes de aceleração c_1 e c_2 são utilizados para variar o quanto os componentes cognitivo e social irão influenciar na mudança da posição das partículas. Os valores de c_1 e c_2 são fixos e determinados inicialmente com base nas características do problema analisado.

Já os números aleatórios U_1^t e U_2^t servem para dar à simulação uma aparência interessante e "realista", de modo a evitar que as partículas assumam uma direção monótona e imutável (KENNEDY e EBERHART, 1995). A cada iteração, são estabelecidos novos valores para tais variáveis.

O *personal best* é a melhor posição assumida por determinada partícula entre todas as posições assumidas até aquela iteração. Já o *global best* é a melhor posição assumida dentre todas as partículas estudadas em todas as iterações realizadas até então (DORIGO et al, 2008). A determinação da melhor posição é realizada com o auxílio de uma função objetivo.

Neste trabalho, foi adotado como função objetivo o *quantization error*, que é uma forma de avaliar a coesão interna dos clusters gerados. Trata-se de uma métrica obtida a partir da soma do quadrado da distância entre o ponto médio do grupo, o centroide, e cada um dos pontos do conjunto. Quanto menor for o valor do erro de quantização, melhor é o resultado da clusterização (ZHU, 2022).

O PSO pode ser aplicado por um determinado número fixo de iterações, até que as posições das partículas não mudem mais do que uma determinada tolerância e/ou até que um determinado valor da função objetivo seja alcançado.

O PSO objetiva que as partículas distribuídas inicialmente de forma aleatória no espaço convirjam para o ponto de mínimo (ou máximo) global. Portanto, ao utilizar-se o k-

means em conjunto com o PSO, busca-se obter melhores resultados no agrupamento dos dados (PATEL et al, 2017).

Existem diferentes formas de se aplicar o k-means em conjunto do PSO. Em um desses métodos, os centroides criados inicialmente de forma aleatória passam por um melhoramento pelo PSO para, só após isso, serem utilizados pelo k-means. Dessa forma, busca-se contornar um dos problemas do algoritmo de clusterização, relacionado à dependência ao centroide inicialmente selecionado (PEDNEKAR, 2019).

Outra maneira de aplicar o k-means junto do PSO é utilizar as técnicas de forma alternada. Ou seja, a cada determinado número de iterações da aplicação do k-means, utiliza-se o PSO para otimizar as posições de uma parte ou de todas as partículas. Dessa forma, busca-se melhorar a velocidade de convergência do algoritmo e a redução do custo computacional exigido (COELHO FILHO et al, 2013).

Além dos métodos citados, pode-se ainda aplicar inicialmente o k-means e depois utilizar o PSO. Isto é, dado um *dataset*, é aplicado inicialmente nos dados o k-means, obtendo assim as informações dos centroides. Esses dados são, então, otimizados com o auxílio do PSO, alcançando os valores finais dos centroides, que são os utilizados na separação dos dados nos clusters (PATEL et al, 2017). Como o k-means apresenta rápida convergência inicial e o PSO possui rápida convergência final, tal maneira de aplicação busca aproveitar o melhor de cada um dos métodos.

Neste trabalho, optou-se por utilizar a última técnica apresentada para aplicar o k-means junto ao PSO. Ou seja, dado o *dataset* analisado, será executado primeiro o k-means, e os resultados desse serão utilizados como entrada para a otimização com o PSO.

3.2.1. Análise das componentes principais (PCA)

A análise das componentes principais ou *Principal Component Analysis* (PCA) é um algoritmo matemático que tem como objetivo reduzir a dimensionalidade dos dados mantendo a maior parte da variação no *dataset* (RINGNÉR, 2008).

O PCA é um procedimento matemático que usa uma transformação ortogonal para converter um grupo de variáveis possivelmente correlacionadas em um conjunto de variáveis linearmente não correlacionadas, que são chamadas de componentes principais. A quantidade de componentes principais é sempre menor ou igual à quantidade de variáveis originais (NEVES, 2016).

Esse método é aplicado em situações onde um dado conjunto de dados apresenta um grande número de variáveis (dimensões), o que dificulta a visualização das amostras e limita a exploração simples dos dados (RINGNÉR, 2008). Em alguns tipos de aplicações, a alta dimensionalidade do *dataset* também pode prejudicar os resultados dos algoritmos.

Como será visto mais adiante, o *dataset* trabalhado apresenta uma série de variáveis. Logo, para facilitar a análise dos dados e melhorar a visualização gráfica dos resultados, o PCA será aplicado durante o processo de clusterização usando o k-means em conjunto com o PSO.

3.2.2. Implementação em *python*

Como exposto na metodologia, os algoritmos utilizados nesta dissertação foram implementados na linguagem de programação *python* e foram executados através da plataforma Google Colab.

Para a clusterização dos dados usando o k-means em conjunto com o PSO, seguindo as informações apresentadas anteriormente, foi utilizado o pseudocódigo exposto a seguir.

Pseudocódigo da clusterização dos dados usando k-means, PSO e PCA.

INÍCIO
\\Dados iniciais
Ler *dataset*
Selecionar variáveis para análise
Definir parâmetros *w*, c_1 e c_2
Definir número de clusters

\\Preparando os dados
Normalizar *dataset*
Aplicar PCA no *dataset* **para redução da dimensionalidade para 2 variáveis**

\\Realização da clusterização
Aplicar k-means no *dataset*
Selecionar valores dos centroides
Aplicar PSO nos centroides
Separar *dataset* **de acordo com os** centroides do PSO
Retornar centroides do PSO, predição do *dataset* **e** *quantization error*
FIM

A seleção das variáveis que serão utilizadas na análise é necessária pois diferentes cenários de estudo podem ser estabelecidos, mudando as variáveis avaliadas em cada situação. Já a realização da normalização dos dados antes da aplicação do PCA é desejável para garantir que os dados sejam modelados corretamente.

CAPÍTULO IV

4. DATASET

O *dataset* que será utilizado neste trabalho é composto por dados de licitações públicas realizadas no Estado da Paraíba. Para isso, foram obtidas informações através do Sistema de Acompanhamento da Gestão dos Recursos da Sociedade (Sagres), que é um importante instrumento do Tribunal de Contas do Estado da Paraíba (TCE/PB).

O portal Sagres Online é uma poderosa ferramenta de controle social no qual as cidades e o Estado da Paraíba são obrigados a alimentar com informações acerca de todos os contratos vigentes, seja com pessoa física ou jurídica, sob pena de punição em caso de descumprimento (TCE/PB, 2010). As licitações que devem ser cadastradas no sistema são todas aquelas realizadas por órgãos paraibanos, independentemente de a verba utilizada na contratação ser de origem estadual, municipal ou federal.

Os dados podem ser obtidos em https://tce.pb.gov.br/servicos/dados-abertos-do-sagres-tce-pb. O acesso às informações ocorreu em junho de 2022. A planilha disponibilizada pelo TCE/PB apresentava cerca de 300 mil dados de licitações que ocorreram no Estado da Paraíba entre 2014 e 2021.

Cada licitação cadastrada no portal Sagres Online apresenta uma série de informações, que são o protocolo da licitação, o número do certame, a modalidade da licitação, o nome e o código do município que organizou a contratação, o código, o nome, o tipo e o tipo de administração do jurisdicionado, o objeto da licitação, o valor estimado para a licitação, o valor licitado, a data e o ano de homologação da licitação, a situação da licitação (fracassada ou não), o nome e o CPF/CNPJ do proponente e a sua proposta, a situação da proposta (vencedora ou perdedora), o estágio processual da licitação, o setor atual do certame e o site que apresenta os dados resumidos da licitação.

Outros dados podem ser obtidos a partir das informações fornecidas na planilha do TCE/PB, como a quantidade de concorrentes em cada certame e o desconto no valor da licitação oferecido por cada proponente.

Com relação às licitações, devido à algumas características delas e às normas brasileiras, foram aplicados alguns filtros aos dados. Inicialmente, foram eliminadas todas as licitações que ocorreram na modalidade Pregão Eletrônico. Foram removidas também as linhas que não apresentavam informações do valor da licitação e/ou do valor da proposta da empresa.

Para todas as licitações, foram removidas aquelas que apresentavam valor de licitação menor do que 15 mil reais. Para os certames que ocorreram até 2019, foram eliminados os dados cujo valor da proposta foi menor do que 15 mil reais. Para as licitações realizadas a partir de 2019, foram excluídas as informações dos certames com proposta menor que 33 mil reais, devido ao Decreto Nº 9.412/2018 que alterou o valor mínimo para despensa de licitação em obras de engenharia (Brasil, 2018).

No campo do munícipio responsável pela licitação, em boa parte dos dados essa informação foi preenchida com “Null”. Analisando mais detalhadamente os dados, verificou-se que essa situação ocorreu em órgãos estaduais, que atuam em mais de um município. Dessa forma, não seria possível atrelar a licitação a apenas uma cidade. Apesar disso, nenhum dado foi eliminado em razão da coluna do município.

Realizando a aplicação de tais filtros, a quantidade de dados foi reduzida para aproximadamente 200 mil linhas. São informações de licitações realizadas para a compra de remédios, para a construção de escolas, para a manutenção de veículos, para a aquisição de combustível etc. Como o objetivo deste trabalho é realizar a análise de cartéis em licitações relacionadas à construção civil, foi aplicado um segundo filtro aos dados, para restringir o tipo do certame.

Para isso, foram utilizadas as informações presentes no campo “Objeto da licitação”. Esse processo apresentou algumas dificuldades relacionadas, sobretudo, ao preenchimento incorreto e/ou incompleto dos dados, ao excesso de erros gramaticais, à falta de padrão nos textos e, também, à falta de informações importantes ou ao excesso de dados irrelevantes.

Devido aos fatores supra mencionados, possivelmente algumas licitações relacionadas à construção civil foram eliminadas indevidamente. Contudo, como após a seleção dos dados foi obtido um conjunto composto por 30 mil linhas de informações de 14 mil licitações, a exclusão de tais dados foi mantida.

As 14 mil licitações restantes possuem mais de 17 mil proponentes vencedores. O número maior de empresas contratadas do que de certames é justificado pelas licitações por lote. Nesse tipo de certame, o objeto de contratação é dividido em vários grupos de modo que mais empresas participem da disputa, aumentando a concorrência e minimizando as chances de problemas por descumprimento de contrato (DA UNIÃO, 2010).

Concorreram nos certames 3.719 empresas diferentes, sendo 331 proponentes com cadastro de pessoa física (CPF) e 3.388 com cadastro nacional de pessoa jurídica (CNPJ). Para manter a privacidade das empresas e das pessoas, cada concorrente recebeu um código aleatório de 4 números para representa-lo. Assim, quando for necessário citar algum proponente, será utilizado tal rótulo.

As licitações selecionadas foram classificadas em 7 categorias, que são “abastecimento”, “material/equipamento”, “obra”, “outros”, “pavimentação”, “projeto” e “reforma”. Na categoria “abastecimento” estão inclusas as licitações relacionadas às obras de esgotamento sanitário e de coleta e distribuição de água.

Estão inclusos no grupo “material/equipamento” os certames que tiveram como objetivo a compra de materiais de construção civil ou o aluguel de equipamentos para usos em serviços da engenharia. Em “outros” foram alocadas as licitações com o objetivo de contratação de pessoal e as que não se enquadravam em nenhum dos demais grupos.

O grupo “pavimentação” contém as informações das licitações que visaram a pavimentação de ruas. Já em “projeto” estão os certames que contrataram empresas para a elaboração de projetos relacionados à engenharia civil. Em “reforma”, estão contemplados os certames de reforma e de ampliação de construções pré-existentes.

Por fim, no conjunto “obra” estão presentes as contratações para construção de prédios. Também foram inclusos nesse grupo os serviços que englobavam mais de uma categoria, ou seja, os certames que envolviam, por exemplo, tarefas de pavimentação e de abastecimento ou de pavimentação e de reforma.

Após a realização da seleção e da classificação das licitações, foi obtida a divisão dos dados em relação ao seu tipo, que pode ser visualizada no gráfico exibido na Figura 4.1.

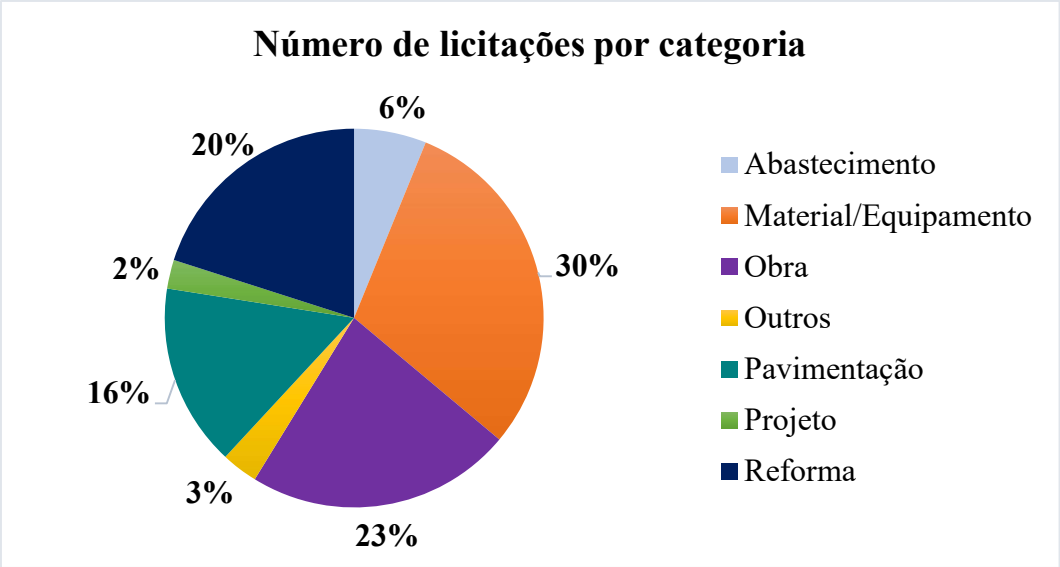


Figura 4.1 – Divisão dos dados segundo o tipo da licitação.

Como pode-se ver, o maior grupo de dados é o "Material/Equipamento", enquanto o menor é o "Projeto". Já analisando a divisão dos dados segundo o ano de ocorrência da licitação, que variou entre 2014 e 2021, obtém-se a divisão apresentada no gráfico mostrado na Figura 4.2.

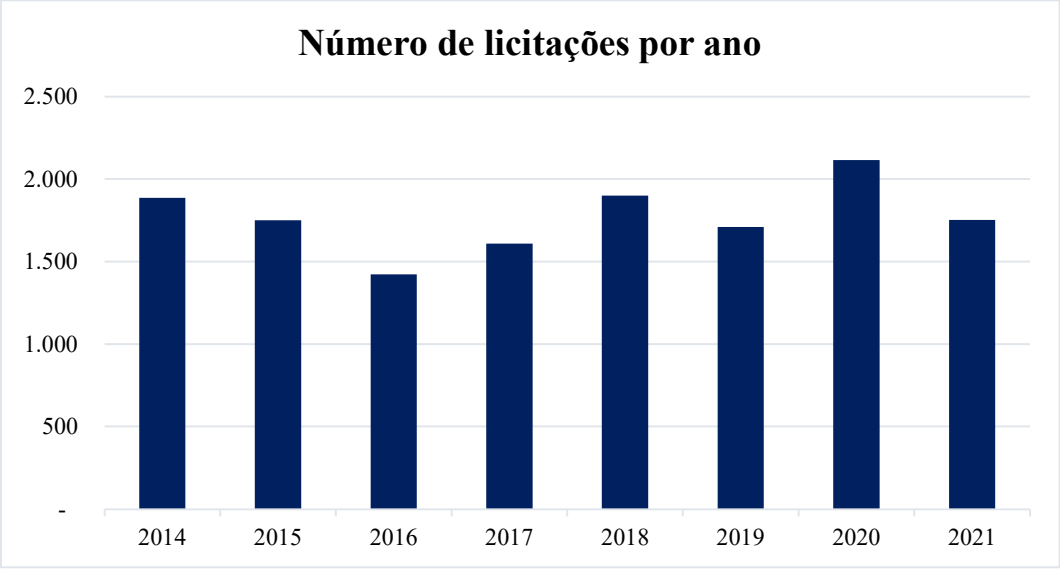


Figura 4.2 – Divisão dos dados segundo o ano da licitação.

Pode-se constatar que o ano em que foram realizadas mais licitações relacionadas à construção civil foi 2020, enquanto o com menos certames foi 2016. As 14 mil licitações movimentaram uma quantia de mais de 10 bilhões de reais ao longo dos 8 anos. No gráfico exibido na Figura 4.3 a seguir é possível ver como esse dinheiro foi gasto nesse período.

Figura 4.3 – Divisão dos valores gastos segundo o ano da licitação.

Salienta-se que os valores apresentados no gráfico da Figura 4.3 são referentes aos preços praticados à época de contratação. Corrigindo as quantias para valores de dezembro de 2021 utilizando o Índice Nacional da Construção Civil (INCC) e considerando o último mês de cada ano, obtém-se a nova distribuição apresentada na Figura 4.4.

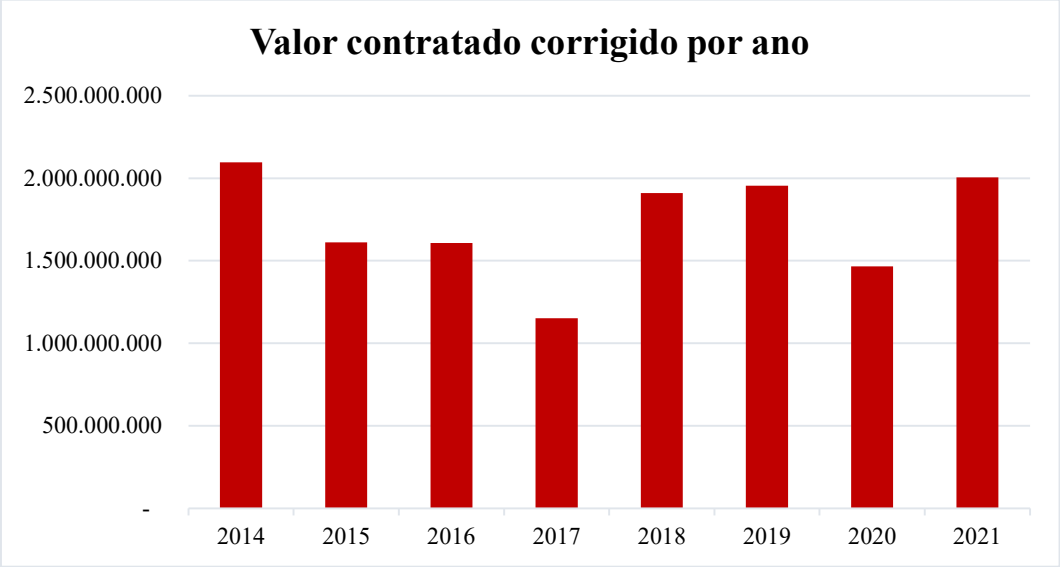


Figura 4.4 – Divisão dos valores atualizados segundo o ano da licitação.

Comparando os valores atualizados, verifica-se que o ano em que foi realizado o maior investimento equivalente na construção civil foi 2014. Já 2017 foi o ano que teve proporcionalmente o menor gasto de verbas públicas. Confrontando a divisão dos dados

segundo o ano da licitação (QTD percentual) com o valor percentual gasto em cada ano com os certames (Valor percentual), obtém-se o gráfico apresentado na Figura 4.5.

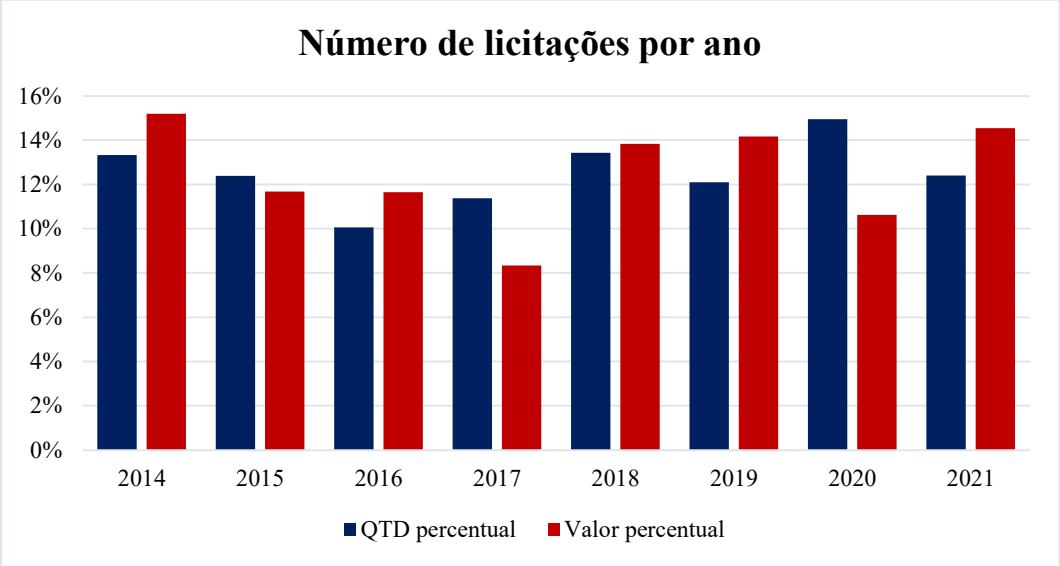


Figura 4.5 – Comparação entre a divisão dos dados e o valor percentual gasto por ano.

Observando a Figura 4.5, nota-se que apesar de em 2017 o número de licitações realizadas ter sido maior do que o de 2016, o valor investido, ao contrário do que seria esperado, diminuiu. Um cenário semelhante a esse também é verificado entre os anos de 2019 e 2020. Apesar do comportamento anormal, não é possível afirmar, apenas analisando os números, que houve algum tipo de erro ou de fraude.

Como comentado, boa parte dos dados (1.760 certames) não apresenta a informação sobre qual município que realizou a licitação. Ignorando esses dados, foi possível também analisar a distribuição das licitações de modo geográfico. Para isso, as informações dos 223 municípios paraibanos foram divididas em quartis.

O primeiro quartil compreende as cidades que realizaram até 36 licitações nos 8 anos estudados. As do segundo quartil fizeram entre 37 e 46 certames, enquanto as do terceiro quartil foram responsáveis por 47 a 64 licitações. As do último quartil, por fim, realizaram mais de 64 certames no período analisado.

Com os municípios divididos nos grupos, foi possível plotar os dados, como pode ser visto no mapa exibido na Figura 4.6.

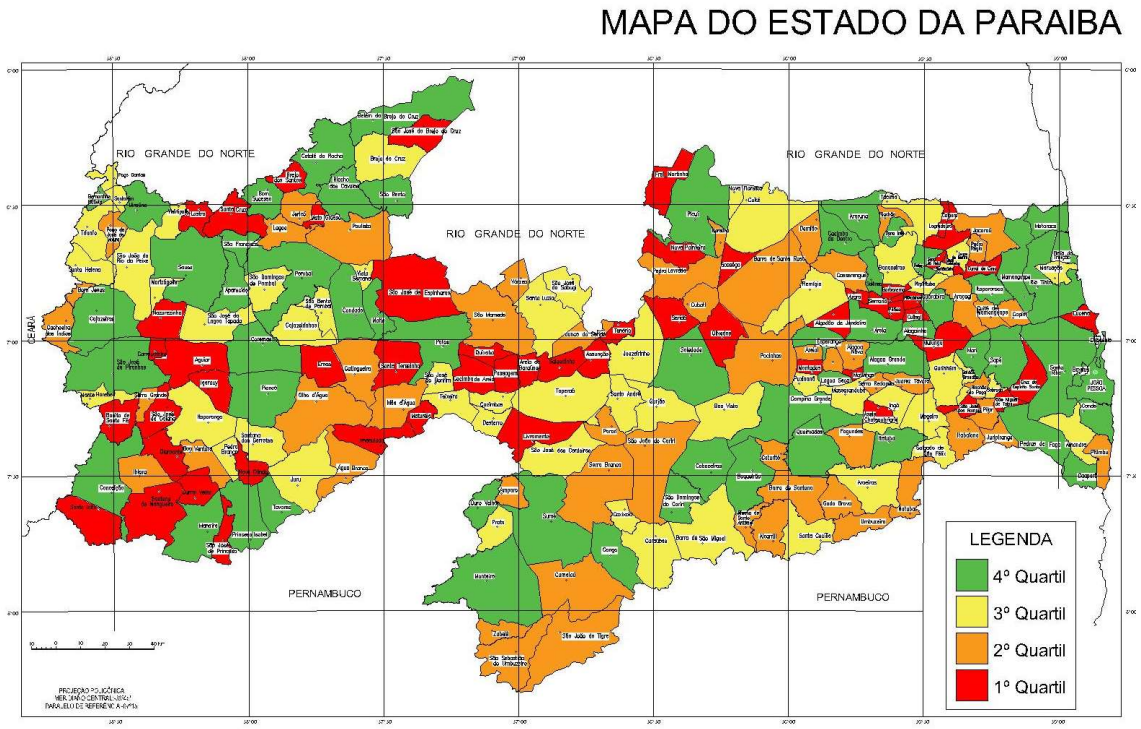


Figura 4.6 – Distribuição das licitações nas cidades paraibanas.

Analisando o mapa da Figura 4.6, nota-se uma grande dispersão dos quartis analisados ao longo do mapa do Estado da Paraíba. Dessa forma, constata-se que a localização geográfica das cidades não está relacionada com a quantidade de licitações de engenharia civil realizadas no período analisado.

4.1. SELEÇÃO DAS VARIÁVEIS

Como comentado, o *dataset* disponibilizado pelo Sagres Online apresenta uma série de informações sobre os dados. Essas variáveis são apresentadas na Tabela 4.1.1 a seguir.

Tabela 4.1.1 – Variáveis disponibilizadas pelo Sagres Online.

Variável	Variável
Protocolo da licitação	Valor licitado
Número da licitação	Data da homologação da licitação
Modalidade da licitação	Ano da homologação da licitação
Nome do munícipio	Situação da licitação
CD gestora	Nome proponente
Jurisdicionado	CPF/CNPJ Proponente
Nome jurisdicionado	Valor proposta
Tipo jurisdicionado	Situação da proposta

Variável
Tipo da administração do jurisdicionado
Esfera jurisdicionado
Objeto da licitação
Valor estimado

Variável
Estágio processual da licitação
Setor atual da licitação
URL

Observando a Tabela 4.1.1, nota-se que algumas das variáveis disponibilizadas referem-se às empresas que participaram dos certames, como o "Nome do proponente" e o "CPF/CNPJ Proponente", e outras estão relacionadas às licitações, como a "Modalidade da licitação" e o "Objeto da licitação".

Como as análises que serão realizadas neste trabalho tanto usando grafos quanto utilizando os algoritmos k-means e PSO buscam identificar relações entre empresas, foram selecionadas apenas as variáveis que estavam relacionadas aos proponentes. Além dessas, a partir dos dados disponibilizados pelo TCE/PB, outros parâmetros também puderam ser obtidos.

A Tabela 4.1.2 a seguir exibe as variáveis que foram consideradas nas análises, o que elas representam e como elas foram obtidas (quando for o caso).

Tabela 4.1.2 – Variáveis selecionadas para utilização nas análises.

Variável	Descrição
CPF/CNPJ Proponente	Documento de identificação da empresa.
Código da empresa	Afim de preservar a identidade das empresas, foi utilizado um código aleatório de 4 dígitos para representar os proponentes.
N. participações	Representa a quantidade de vezes que a empresa participou de licitações durante o período analisado.
N. vitórias	Representa a quantidade de vezes que a empresa venceu uma licitação durante o período analisado.
Taxa vitórias	Relação entre o número de vitórias e o número de participações em licitações da empresa analisada.
Total proposto	Soma de todos os valores licitados pela empresa analisada.
Total vencido	Montante dos valores licitados das licitações que a empresa analisada venceu.
Percentual	Relação entre o total vencido e o total proposto pela empresa.
Qtd de anos	Refere-se à quantidade de anos que a empresa participou de licitações.
N. licitações por ano	Relação entre o número de participações e a quantidade de anos.

Variável	Descrição
N. cidades	Quantidade de cidades em que a empresa analisada participou de processos licitatórios.
N. cidades vitoriosas	Quantidade de cidades em que a empresa analisada venceu processos licitatórios.

4.2. CORREÇÃO DOS DADOS

Como será visto mais adiante, para a realização das análises dos dados utilizando o k-means e o PSO, foi necessário utilizar as variáveis relacionadas ao custo das licitações. Entretanto, os dados fornecidos pelo TCE/PB apresentavam algumas inconsistências, sobretudo relacionado ao valor estimado das licitações e, consequentemente, ao desconto aplicado pelos proponentes.

O desconto de uma licitação é dado pela relação entre o valor proposto pela empresa e o valor estimado para essa licitação. Teoricamente, o valor do desconto deve ser menor que 100% e maior ou igual à 0%, pois nem a empresa realizará o serviço de graça, nem receberá um valor maior do que o máximo estipulado. Entretanto, valores de descontos muito elevados também não são esperados, visto que os valores máximos determinados para as licitações não se distanciam muito dos valores reais praticados pelo mercado.

As inconsistências verificadas nos custos estimados das licitações podem ter ocorrido devido à erros no preenchimento das informações no Sagres, seja por falhas de digitação, seja por problemas na interpretação de quais dados que deveriam ser corretamente fornecidos (SIGNOR et al, 2020).

Para exemplificar os tipos de erros identificados, na Tabela 4.2.1 são apresentados três registros do portal Sagres Online com erros no preenchimento das informações.

Tabela 4.2.1 – Exemplos de inconsistências nos dados do portal Sagres Online.

Nº da licitação	Modalidade da licitação	Jurisdicionado	Valor estimado da licitação	Valor da proposta
00001/ 2014	Tomada de Preços	Prefeitura Municipal de Itatuba	R$ 408.397,04	R$ 40.739,63
20654/ 2015	Tomada de Preços	Secretaria da Administração de Campina Grande	R$ 552.510,62	R$ 1.035.130,22
33043/ 2018	Concorrência	Secretaria Municipal de Planejamento de João Pessoa	R$ 11.798.840,70	R$ 887.709,00

A licitação realizada pela Prefeitura Municipal de Itatuba teve como objetivo a contratação de uma empresa para a construção de uma Unidade Básica de Saúde da Família. Analisando a documentação do certame fornecido pelo TCE/PB, nota-se que o contrato assinado entre a empresa vencedora e o órgão que organizou a licitação foi de R$ 407.739,63 (quatrocentos e sete mil, setecentos e trinta e nove reais e sessenta e três centavos).

Assim, constata-se que ocorreu um erro no preenchimento do valor da proposta da licitação no portal Sagres Online, o que gerou um valor de desconto inconsistente. De acordo com os dados da Tabela 4.2.1, o desconto oferecido pela empresa seria de 90,02%, quando na verdade a redução foi de apenas 0,16%.

Já o certame organizado pela Secretaria de Administração de Campina Grande objetivou a contratação de uma empresa para a construção de unidades sanitárias em escolas municipais. Analisando os valores exibidos na Tabela 4.2.1 e os dados dos documentos da licitação fornecidos pelo TCE/PB, nota-se que o valor estimado da licitação foi preenchido de forma errada.

O valor correto do custo estimado para a Tomada de Preços Nº 20654/2015 é de R$ 1.052.510,62 (um milhão, cinquenta e dois mil, quinhentos e dez reais e sessenta e dois centavos), o que gera um valor de desconto de 1,65%, valor mais coerente do que o obtido usando os dados da Tabela 4.2.1.

Por fim, o certame realizado pela Secretaria Municipal de Planejamento de João Pessoa, que foi realizado como uma licitação de lote, buscou a contratação de empresas para a realização de reformas em escolas públicas. Observando os valores preenchidos para o valor estimado da licitação e para o valor da proposta e levando em conta que se trata de uma licitação de lote, verifica-se que nesse caso o erro que ocorreu foi na interpretação de quais dados deveriam ser preenchidos no sistema do Sagres.

O registro destacado na Tabela 4.2.1 é referente a uma empresa que participou de apenas uma parte dos lotes da licitação 33043/2018. Porém, o valor estimado informado refere-se ao total do certame. Dessa forma, o valor calculado do desconto (92,48%) não representa a realidade da licitação.

Além desses exemplos, muitos outros dados fornecidos pelo TCE/PB também apresentam inconsistências. Logo, como para a clusterização dos dados foi preciso utilizar as informações relacionadas aos custos das licitações, os dados brutos obtidos do Sagres Online deveriam passar por um processo de verificação e correção de erros.

Porém, como exposto na seção anterior, após a filtragem dos dados oriundos do TCE/PB, restaram cerca de 30 mil linhas de informações referentes à 14 mil processos licitatórios. Tendo em vista que o procedimento de verificação e correção dos dados é realizado de forma manual, a revisão de tal conjunto de informações seria extremamente trabalhosa e demandaria bastante tempo.

Para contornar esse empecilho, considerando que o objetivo deste trabalho é sugerir metodologias para auxiliar o processo investigativo de licitações, foram selecionadas 1.595 das 14 mil licitações para serem corrigidas e agrupadas. Dessa forma, o método proposto pôde ser demonstrado e sua eficácia comprovada sem maiores prejuízos.

As 1.595 licitações escolhidas foram selecionadas segundo os seguintes critérios: limitação às licitações da modalidade “Tomada de Preços”, restrição às licitações dos tipos “Obra” e “Reforma”, exclusão das licitações com apenas 1 concorrente, remoção das licitações de lote e limitação do número de concorrentes do certame para até 8 empresas. Após a filtragem dos dados, restaram 728 empresas para análise.

4.3. EMPRESAS INVESTIGADAS

Para avaliar os modelos gerados neste trabalho, um dado que será útil é o das empresas que já sofreram algum tipo de investigação acerca de fraudes em licitações públicas. Com essa informação, será possível identificar, por exemplo, possíveis proponentes que ainda não foram investigados, mas que podem estar envolvidos em irregularidades.

Para a obtenção desses dados, foi utilizada a página do Tribunal de Contas do Estado da Paraíba que permite consultar processos que tramitam ou tramitaram no órgão. As pesquisas foram realizadas utilizando os números de CNPJ das empresas e os nomes das pessoas físicas.

Os processos disponíveis na base de dados do TCE/PB são divididos em várias categorias. Os proponentes que apresentavam processos relacionados a irregularidades em licitações e a solicitações de prestação de contas foram sinalizados como concorrentes suspeitos.

Cabe ressaltar que os dados consultados não são determinísticos, isto é, se uma empresa apresenta um processo, não significa necessariamente que ela seja culpada. Contudo, um processo investigativo só é instaurado quando há suspeitas e/ou sinais de

irregularidades. Por isso que, mesmo que o proponente seja inocente, ele será sinalizado como concorrente suspeito.

Por outro lado, caso não haja um processo contra o proponente, não significa necessariamente que ele seja inocente. Em investigações, empresas culpadas podem ser inocentadas devido à falta de provas ou à detalhes técnicos. Além disso, como a justiça brasileira é extremamente lenta, pode demorar anos até que uma licitação que apresenta irregularidades comece a ser examinada.

Outro ponto a ser analisado refere-se à origem dos dados consultados. O TCE/PB disponibiliza informações referentes apenas ao estado da Paraíba. Logo, caso uma empresa tenha atuado, também, em outro local e tenha sido investigada lá, esses dados não estarão disponíveis.

Para que essas informações fossem incluídas neste estudo, bases de dados de todos os Tribunais de Contas de todas as unidades da federação deveriam ser consultadas, além do Tribunal de Contas da União (TCU), da Controladoria Geral da União (CGU) e dos demais órgãos de controle. Por não existir uma base de dados unificada, a consulta a tais entidades se torna inviável.

Portanto, foram utilizadas para a avaliação dos modelos apenas as informações obtidas da base do TCE/PB. A análise será realizada apenas de forma exemplificativa, ou seja, será mostrado apenas como os resultados obtidos podem ser interpretados e analisados.

A partir das consultas realizadas, dos 3.719 proponentes, 202 (5,43%) foram classificados como suspeitos. Fazem parte desse total 125 empresas, representando 3,69% do total desse grupo, e 77 pessoas físicas, correspondendo à 23,26% do conjunto. Para o *dataset* reduzido, das 728 empresas estudadas, 66 foram identificadas como suspeitas, representando 9,07% do total. Esses dados serão utilizados na análise dos resultados obtidos.

CAPÍTULO V

5. RESULTADOS

Como comentado anteriormente, neste trabalho serão empregados dois métodos diferentes: a utilização de grafos e a clusterização utilizando k-means com PSO. Os resultados alcançados com cada uma dessas técnicas podem ser visualizados a seguir nas seções 5.1 e 5.2. Na seção 5.3, é apresentado o resultado final desta dissertação, obtido a partir da combinação dos produtos de tais técnicas.

5.1. GRAFOS

Para aplicar a teoria dos grafos para a resolução do problema proposto, foi utilizado o *dataset* obtido a partir da filtragem e do pré-processamento dos dados fornecidos pelo Sagres Online.

Os grafos exibidos nesta dissertação foram elaborados com o auxílio da linguagem de programação *python*, por meio da biblioteca NetworkX, que permite a exploração e a análise de grafos (HAGBERG et al, 2008).

Logo, para a geração das representações gráficas que serão utilizadas neste trabalho, as empresas que participaram dos processos licitatórios foram escolhidas para serem os nós. Dessa forma, existirá uma aresta entre dois nós A e B caso essas empresas tenham concorrido juntas em algum processo licitatório.

Às arestas foi atribuído um peso que corresponde à quantidade de licitações que os nós (as empresas) participaram juntos. Já aos nós, foi atribuído um peso relativo à quantidade de participações que a determinada empresa realizou nos certames analisados durante o período de tempo estudado.

Assim, utilizando todas as informações do *dataset* após a realização do pré-processamento, foi possível gerar o grafo que representa todas as relações entre as empresas analisadas. Esse resultado pode ser visto na Figura 5.1.1 a seguir.

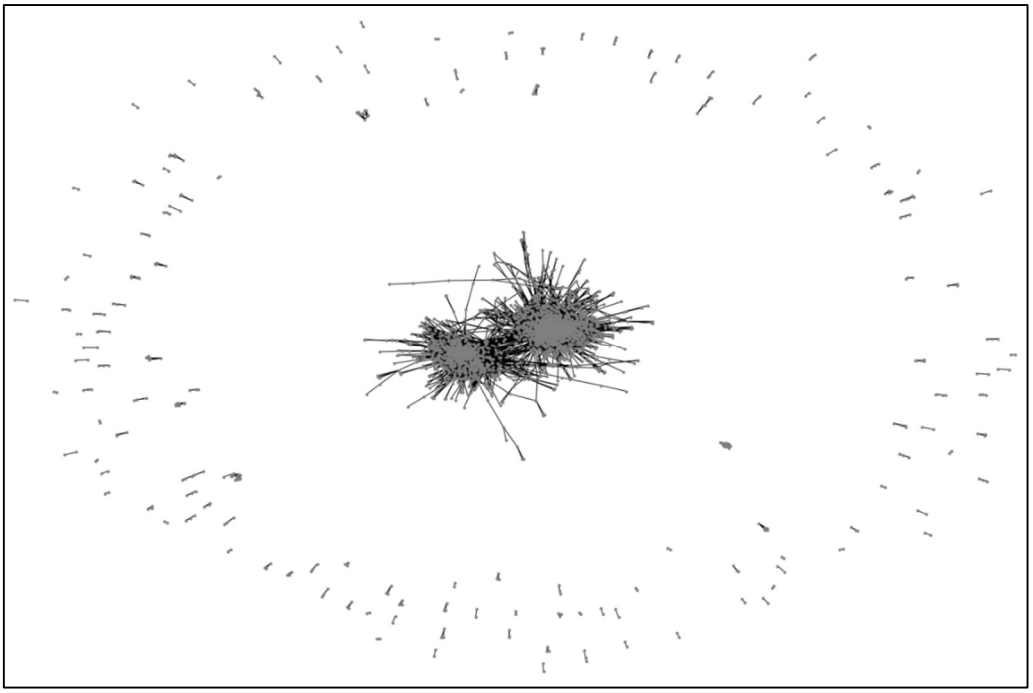


Figura 5.1.1 – Grafo construído utilizando todo o *dataset* analisado.

O grafo exibido na Figura 5.1.1, possui 3.059 nós (empresas) e 24.661 arestas (relações entre os proponentes). Ele é formado por 139 subgrafos, sendo que o maior deles apresenta 2.624 nós. Na Figura 5.1.2 é possível visualizar esse subgrafo com maior detalhamento.

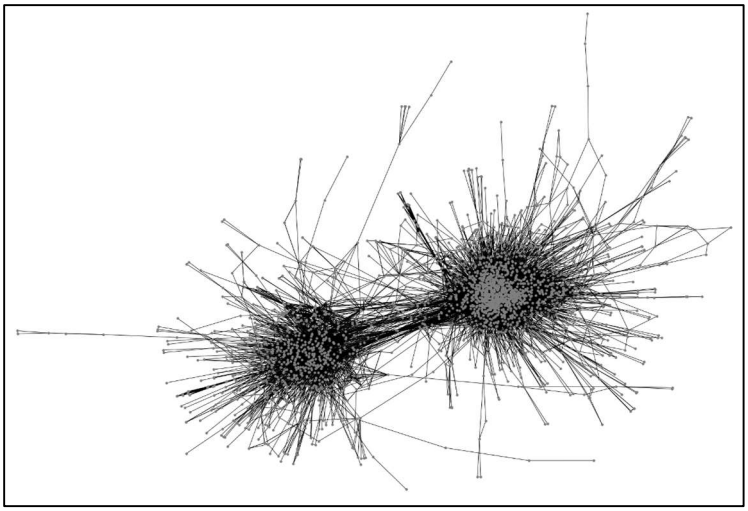


Figura 5.1.2 – Destaque para o maior subgrafo do Grafo apresentado na Figura 5.1.1.

Observando tanto o grafo exibido na Figura 5.1.1 quanto o subgrafo mostrado na Figura 5.1.2, verifica-se que como a quantidade de nós e de arestas é bastante elevado, torna-se difícil visualizar os dados e tirar conclusões.

Para melhorar a análise das informações, é possível aplicar filtros no *dataset* para selecionar apenas os nós que apresentam alguma característica ou para remover as arestas que não cumprem determinado requisito.

Vale salientar que um nó só continuará fazendo parte do grafo caso seu grau seja pelo menos igual a 1, ou seja, caso haja pelo menos uma aresta ligando o nó à outra empresa qualquer. Além disso, dados dois nós A e B conectados por uma aresta C, caso um ou os dois nós sejam removidos devido aos critérios aplicados, a aresta C também será eliminada.

Assim, nas seções 5.1.1, 5.1.2 e 5.1.3 serão apresentados exemplos de filtros que podem ser aplicados nos dados para melhorar a visualização e análise dos grafos. A Tabela 5.1.1 apresenta os filtros aplicados e os principais resultados obtidos.

Tabela 5.1.1 – Filtros aplicados para a geração e análise dos grafos.

Seção	Filtro	Resultado
5.1.1	Limitação do número de participações das empresas	Seleção dos nós que serão exibidos nos grafos
5.1.2	Limitação do número de relações entre as empresas	Seleção das arestas que serão exibidas nos grafos
5.1.3	Análise apenas das empresas investigadas	Análise das relações mantidas entre as empresas investigadas

5.1.1. Aplicação de filtro: limitação do número de participações das empresas

Afim de melhorar a visualização e o tratamento dos dados, foi aplicado inicialmente um filtro com o intuito de limitar os nós do grafo original. Assim, permaneceriam na análise apenas as empresas que apresentassem um determinado número mínimo de participações nos certames examinados.

Usando tal filtro e selecionando apenas as empresas que tiveram mais de 10 participações nos certames durante os 8 anos de estudo, obtém-se o grafo mostrado na Figura 5.1.1.1.

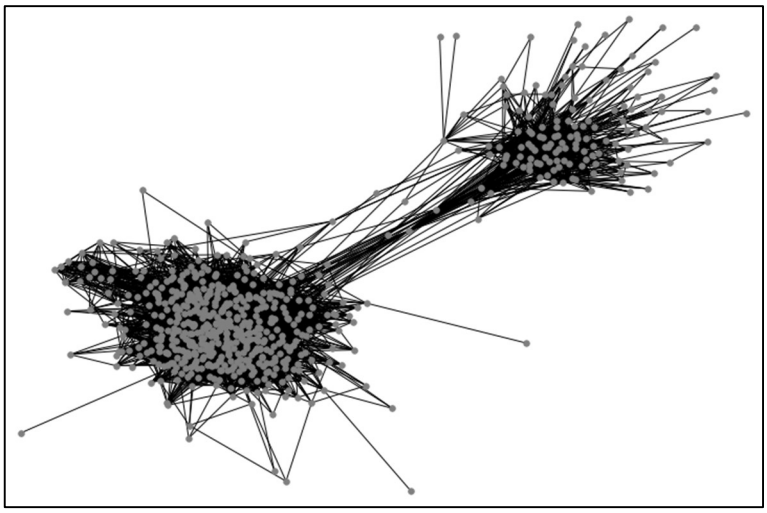


Figura 5.1.1.1 – Grafo das empresas com mais de 10 participações em licitações.

O grafo apresentado na Figura 5.1.1.1 possui 649 nós e 19.153 arestas. Analisando a imagem, nota-se que a aplicação do filtro para seleção apenas das empresas com mais de 10 participações em certames excluiu da análise as empresas que se localizavam nos 138 subgrafos que rodeavam a estrutura central da Figura 5.1.1, que também teve seu tamanho reduzido, diminuindo de 2.624 para 649 nós.

Comparando os grafos das Figura 5.1.1.1 e 5.1.1, verifica-se que apesar de o filtro aplicado ter estabelecido critérios relativamente pequenos para a limitação dos dados, boa parte das informações foi excluída. Essa redução exemplifica a importância da aplicação de filtros nos dados, visto que boa parte das informações presentes no *dataset* original não possuía grande relevância para as análises a serem realizadas.

Utilizando o grafo gerado considerando apenas as empresas que concorreram em pelo menos 10 licitações e plotando as informações acerca das empresas que já passaram por um processo investigativo em algum momento, obtém-se o resultado mostrado na Figura 5.1.1.2. Os nós coloridos em vermelho correspondem às empresas que já foram investigadas, enquanto os nós verdes compreendem empresas não investigadas pelo TCE/PB.

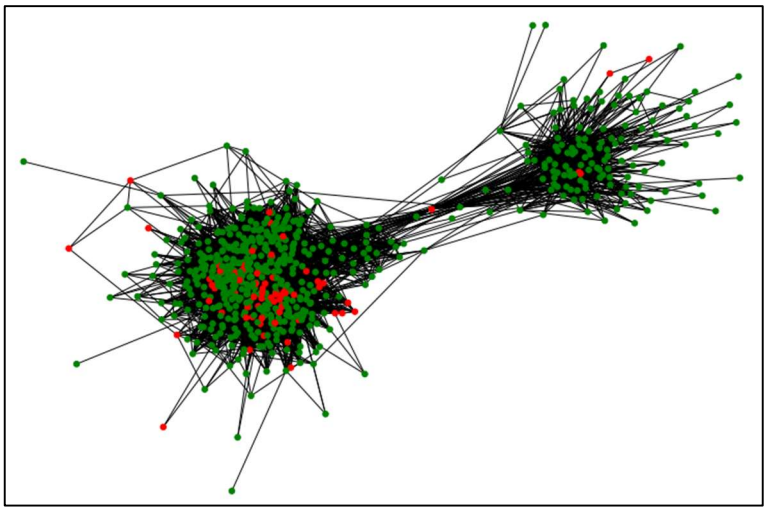


Figura 5.1.1.2 – Grafo das empresas com mais de 10 participações em licitações colorido com base nas informações das empresas investigadas.

Analisando o grafo mostrado na Figura 5.1.1.2, verifica-se que, devido ao ainda elevado tamanho do *dataset* trabalhado, não é possível estabelecer nenhum critério que interrelacione as empresas que já passaram por um processo investigativo (ou que deveriam ser averiguadas).

Retirando a clique máxima do grafo mostrado na Figura 5.1.1.1, obtém-se o subgrafo apresentado na Figura 5.1.1.3. Nas cliques apresentadas neste trabalho, os nós dos subgrafos coloridos de verde referem-se à empresas que nunca foram investigadas pelo TCE/PB e os pintados de vermelho correspondem à proponentes que já foram alvo de investigação. Além disso, o tamanho de exibição dos nós é diretamente proporcional à taxa de vitórias das empresas.

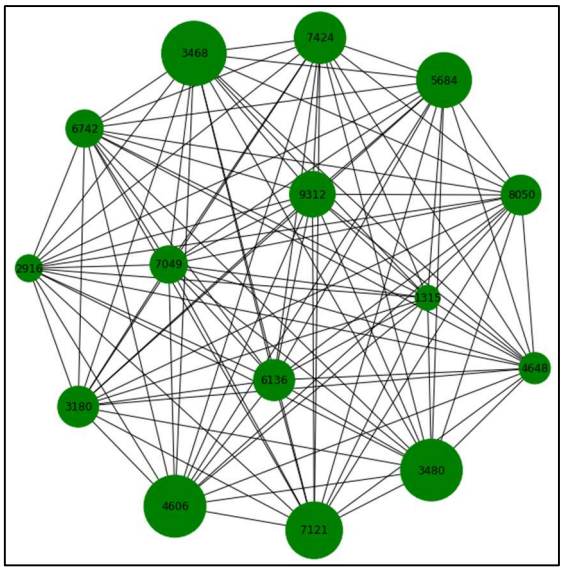


Figura 5.1.1.3 – Clique máxima do grafo da Figura 5.1.1.1.

A clique máxima exibida na Figura 5.1.1.3 apresenta 15 nós. Ela representa empresas que possuem, cada uma, pelo menos 10 participações em licitações durante os 8 anos de estudo e que apresentam um elevado grau de relação entre si. Cada uma das 15 empresas já participou de pelo menos um certame com cada uma das outras 14 companhias.

Participam da clique máxima mostrada na Figura 5.1.1.3 as empresas cujo código de identificação são: 3468, 6742, 2916, 3180, 4606, 7121, 3480, 4648, 8050, 5684, 7424, 9312, 7049, 6136 e 1315. Nenhuma dessas empresas já foi alvo de um processo investigativo movido pelo TCE/PB.

Uma situação que pode ocorrer durante a geração de uma clique máxima é a de uma determinada empresa manter relação com um grande número dos nós do subgrafo, mas não com todos, e, por isso, não ser inclusa na clique máxima. Porém, mesmo não se relacionando com todos os nós, a empresa em questão é extremamente relevante para a investigação de possíveis fraudes em licitações.

Logo, visando incluir tais empresas que não fazem parte da clique máxima, mas que apresentam um elevado grau de relação com o grupo analisado, a clique máxima obtida foi expandida para a obtenção de uma clique quase máxima.

Assim, ampliando a clique máxima apresentada na Figura 5.1.1.3, obteve-se a clique quase máxima mostrada na Figura 5.1.1.4. Para isso, foram adicionados 13 nós, obtendo-se uma densidade de 0,97.

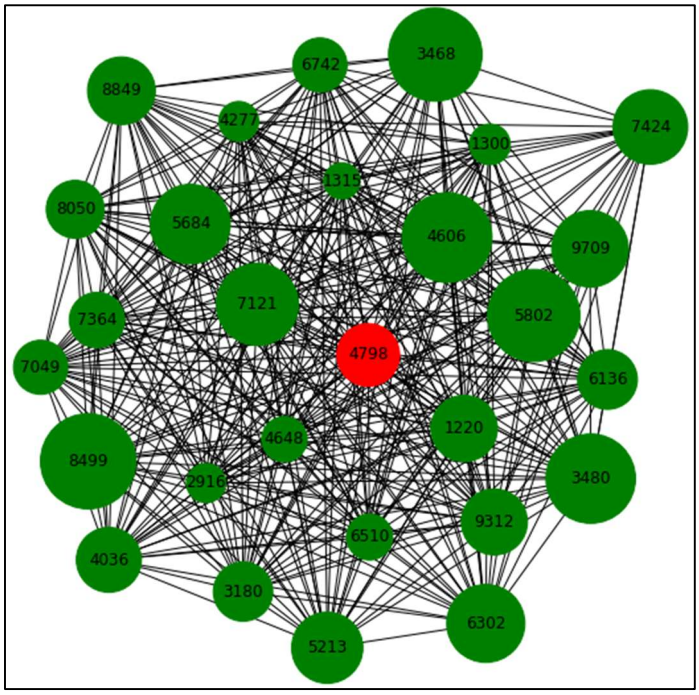


Figura 5.1.1.4 – Clique quase máxima do grafo da Figura 5.1.1.1.

Os nós adicionados para a formação da clique quase máxima apresentada foram: 8849, 1300, 9709, 5802, 6302, 5213, 6510, 4036, 8499, 4798, 1220, 7364 e 4277. Observando a Figura 5.1.1.4, nota-se que se trata de um grafo bastante denso, composto ao todo por 28 nós. Verifica-se a presença de um nó vermelho (código 4798), o que indica que tal empresa já foi alvo de um processo de investigação.

Aplicando novamente o mesmo filtro no conjunto de dados, porém selecionando dessa vez apenas as empresas que tiveram mais de 20 participações nos processos licitatórios durante o período de tempo estudado, obtêm-se as relações exibidas na Figura 5.1.1.5.

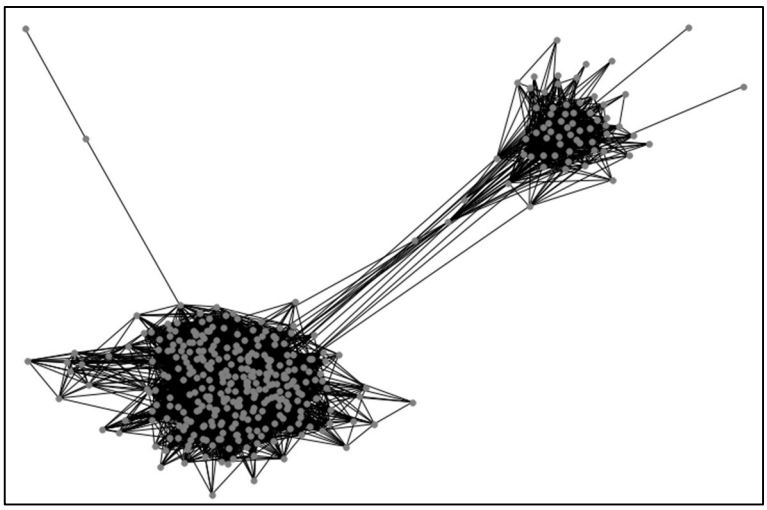


Figura 5.1.1.5 – Grafo das empresas com mais de 20 participações em licitações.

O grafo apresentado na Figura 5.1.1.5 possui 370 nós e 8.418 arestas. Nota-se, comparando tal grafo com o exposto na Figura 5.1.1.1, que o formato do desenho foi mantido. Em ambos os cenários, o grafo apresentou dois grandes grupos de dados conectados entre si por algumas ligações. A filtragem do *dataset* não modificou tal formato, apenas diminuiu a concentração de nós nos grupos.

Comparando as quantidades dos nós e das arestas dos grafos gerados considerando apenas empresas com mais de 10 participações e com mais de 20 presenças em certames, constata-se que o número de nós decaiu 42,99% (de 649 para 370) enquanto a quantidade de arestas foi reduzida em 56,04% (de 19.153 para 8.418).

Colorindo o grafo das empresas com mais de 20 participações de acordo com a informação das companhias já investigadas pelo TCE/PB, obteve-se o resultado exposto na Figura 5.1.1.6.

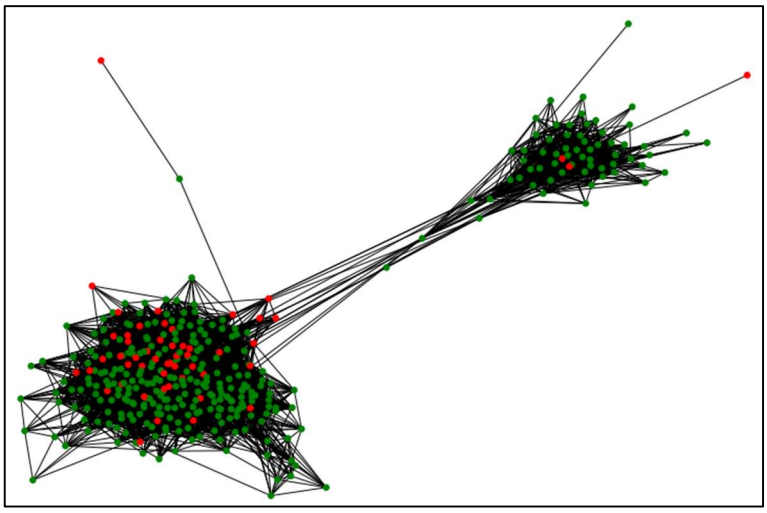


Figura 5.1.1.6 – Grafo das empresas com mais de 20 participações em licitações colorido com base nas informações das empresas investigadas.

Novamente, não é possível estabelecer nenhum critério de divisão das empresas investigadas com base em sua localização no grafo da Figura 5.1.1.6. Destaca-se que existem nós no grafo que estão "escondidos" embaixo da aglomeração de arestas e de nós e, por isso, não são visíveis.

O algoritmo empregado para a seleção da clique máxima identificou um subgrafo completo de 20 nós no grafo da Figura 5.1.1.5. Fazem parte desse subgrafo os nós 6510, 5684, 7121, 7049, 1300, 9312, 7424, 3180, 2916, 3468, 3480, 9709, 8050, 6136, 6742, 1315, 5802, 6302, 4606 e 4648. Essa clique pode ser visualizada na Figura 5.1.1.7 disposta a seguir.

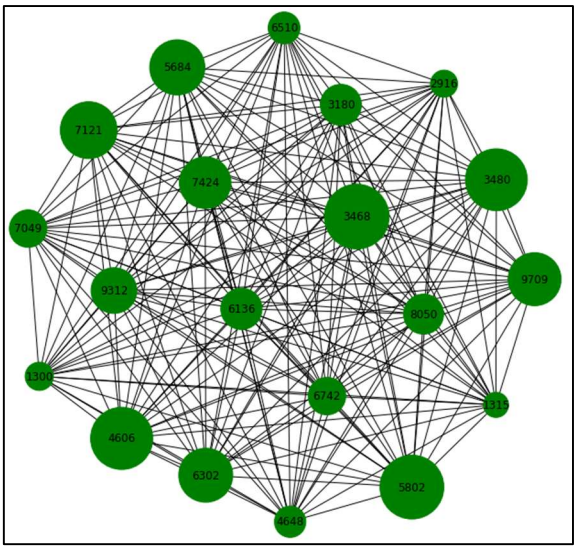


Figura 5.1.1.7 – Clique máxima do grafo da Figura 5.7.

Observando a clique máxima exibida na Figura 5.1.1.7, verifica-se que todos os nós que a compõem são verdes, isto é, a clique é formada em sua totalidade por empresas que nunca foram investigadas pelo TCE/PB.

Para ampliar a clique máxima e, com isso, produzir uma clique quase máxima para o cenário analisado, foram adicionados 10 nós ao grafo mostrado na Figura 5.1.1.7. A densidade do novo subgrafo foi de 0,96.

Os nós incluídos na clique máxima foram os das empresas 5213, 1220, 8499, 8627, 7624, 4798, 7364, 8849, 4036 e 4277. O resultado dessa operação pode ser visualizado na Figura 5.1.1.8.

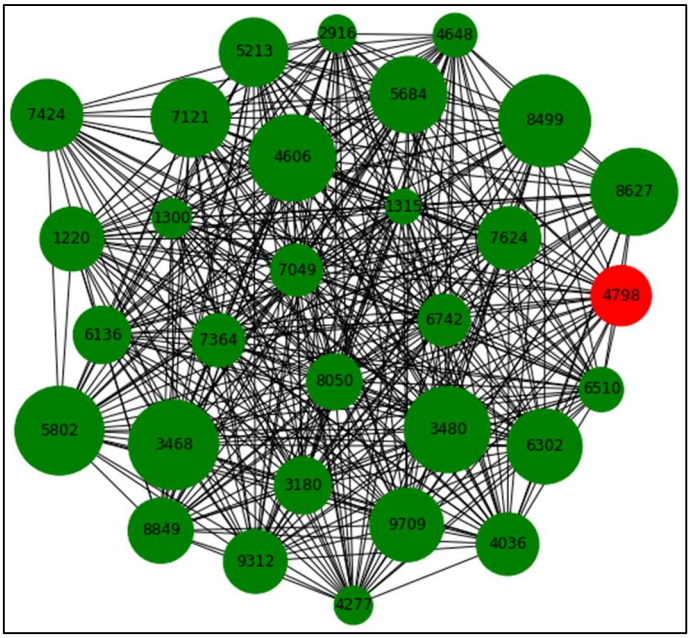


Figura 5.1.1.8 – Clique quase máxima do grafo da Figura 5.1.1.5.

Analisando a clique quase máxima mostrada na Figura 5.1.1.8, verifica-se que um nó vermelho foi adicionado à rede estudada. A presença dessa empresa, aliada a existência de outros fatores suspeitos, pode caracterizar um cenário favorável à instauração de um processo investigativo.

Por fim, aplicando uma última vez o filtro proposto, foram selecionadas apenas os dados de licitações de empresas que concorreram em certames relacionados à construção civil e que participaram de pelo menos 50 processos licitatórios no estado da Paraíba entre os anos de 2014 e 2021. Usando tais informações, foi possível gerar o grafo exposto na Figura 5.1.1.9.

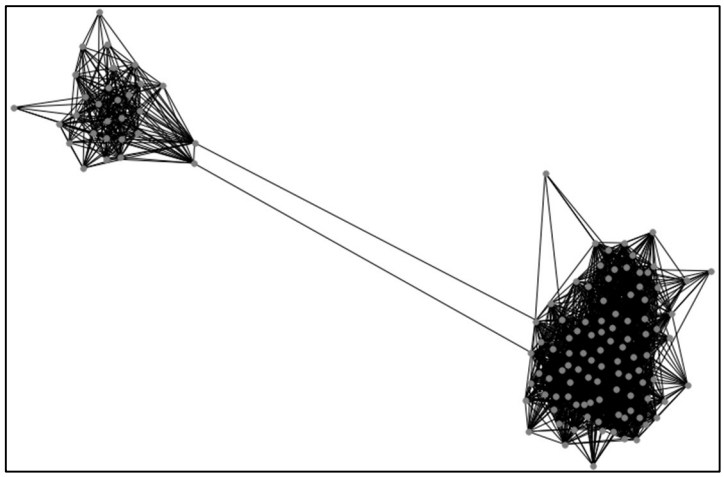


Figura 5.1.1.9 – Grafo das empresas com mais de 50 participações em licitações.

Apesar da grande densidade apresentada no grafo da Figura 5.1.1.9, essa estrutura é formada por apenas 131 nós, conectados por 5.322 arestas. Assim como evidenciado anteriormente, no grafo apresentado também foi verificado a formação de dois grandes grupos de dados.

Com relação à quantidade de nós e de arestas, comparando os grafos produzidos ao reduzir-se o *dataset* para apenas empresas com mais de 20 participações e para companhias com mais de 50 concorrências, nota-se que enquanto o número de nós de uma estrutura para outra reduziu 64,59% (370 para 131), a quantidade de arestas decaiu apenas 36,78% (8.418 para 5.322). Isso evidencia que o grupo com mais participações é também mais denso.

Para avaliar a existência de algum padrão nos dados que pudesse ser relacionado às informações das empresas investigadas, os nós do grafo anterior foram coloridos conforme mostrado na Figura 5.1.1.10.

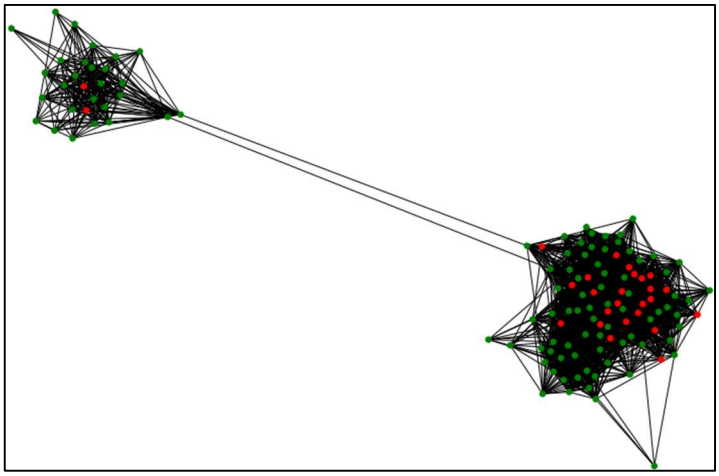


Figura 5.1.1.10 – Grafo das empresas com mais de 50 participações em licitações colorido com base nas informações das empresas investigadas.

Observando o grafo mostrado na Figura 5.1.1.10, pode-se notar uma maior concentração de nós vermelhos no grupo de dados da parte inferior direita. Contudo, verifica-se também que esse conjunto de dados apresenta um tamanho consideravelmente maior do que o outro.

Assim, apesar de um conjunto ter apresentado mais nós vermelhos do que o outro, devido aos diferentes tamanhos dos grupos formados, não é possível determinar nenhuma característica específica de localização dos nós das empresas que as estabeleçam como mais suspeitas do que as demais.

Para analisar mais detalhadamente o grafo gerado, foi identificada uma clique máxima para a estrutura mostrada na Figura 5.1.1.9. Esse subgrafo pode ser visualizado na Figura 5.1.1.11.

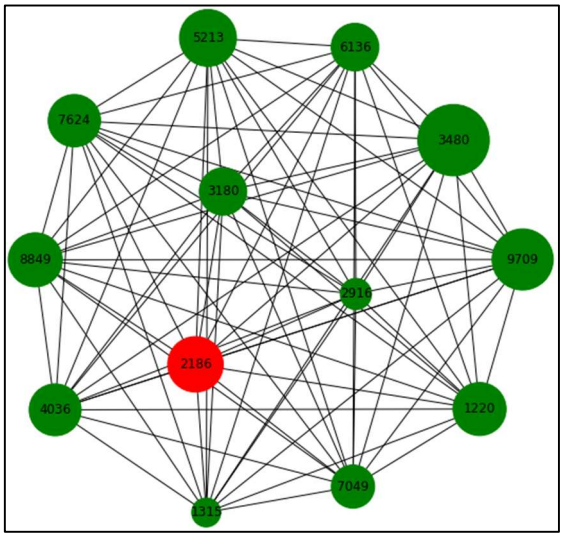


Figura 5.1.1.11 – Clique máxima do grafo da Figura 5.1.1.9.

A clique identificada pelo algoritmo é constituída por 13 nós, que representam as empresas 5213, 6136, 3480, 9709, 2916, 3180, 7624, 8849, 2186, 1315, 4036, 7049 e 1220. Dessa forma, tem-se que cada empresa do subgrafo destacado participou de pelo uma licitação com cada uma das organizações destacadas.

Analisando os 13 nós que formam a clique máxima mostrada na Figura 5.1.1.11, verifica-se que um deles é vermelho. A empresa cujo código de identificação é o 2186 foi alvo, em algum momento, de um processo investigativo movido pelo TCE/PB. A sua presença na clique, aliada a outros fatores, pode indicar uma necessidade de que as outras empresas do subgrafo também sejam investigadas.

Prosseguindo com a análise do cenário de filtragem para apenas empresas com mais de 50 participações, utilizando a clique máxima da Figura 5.1.1.11, foi produzida uma clique quase máxima, a partir da adição de 8 nós a tal subgrafo. O resultado dessa operação pode ser visto na Figura 5.1.1.12.

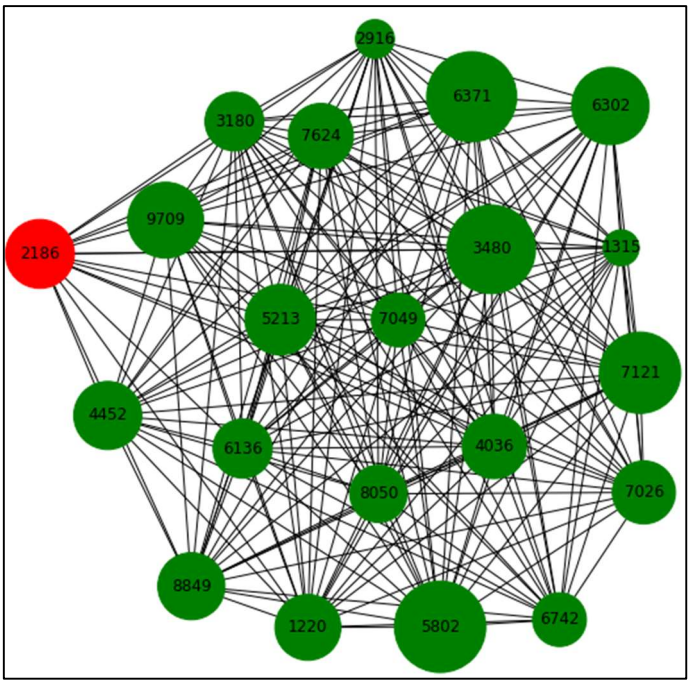


Figura 5.1.1.12 – Clique quase máxima do grafo da Figura 5.11.

A clique quase máxima exibida na Figura 5.1.1.12 apresenta densidade igual à 0,95. Foram adicionados os nós 6371, 6302, 7121, 7026, 8050, 4452, 5802 e 6742 à clique máxima. Todos os nós inclusos pertencem ao grupo das empresas que nunca foram investigadas pelo TCE/PB.

Como foi mostrado nos exemplos apresentados, a limitação dos dados com base no número de participação das empresas em licitações durante o período de tempo analisado é um filtro que pode ser aplicado no *dataset* para melhorar a visualização dos dados e permitir sua análise mais detalhada e assertiva.

5.1.2. Aplicação de filtro: limitação do número de relações entre as empresas

No exemplo anterior, foi sugerido um possível filtro para aplicação nos nós dos grafos. Já nessa seção, será apresentado outro tipo de seleção que também pode ser utilizado no *dataset*, porém para definir as arestas que serão exibidas.

Para isso, serão escolhidas apenas as ligações que apresentam um determinado valor do peso atribuído anteriormente, que relaciona a quantidade de licitações que as empresas adjacentes participaram juntas. Na prática, a aplicação desse filtro significa escolher analisar apenas as empresas que possuem um valor mínimo de relação entre si.

Logo, para exemplificar a utilização do filtro proposto, o *dataset* pré-processado foi filtrado e apenas as arestas com grau maior ou igual à 2 foram mantidas. Ou seja, foram descartados os dados relativos às empresas que concorreram juntas em licitações apenas uma vez. O resultado da aplicação desse filtro pode ser visto na Figura 5.1.2.1.

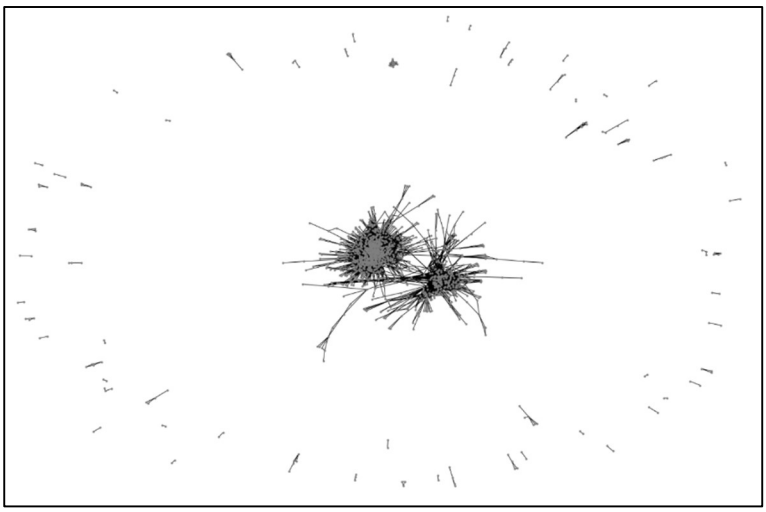


Figura 5.1.2.1 – Grafo das empresas com arestas com peso maior igual à 2.

O grafo apresentado na Figura 5.1.2.1 possui 1.419 nós e 8.581 arestas. Ao contrário do resultado exibido pelo filtro anterior, a filtragem realizada nesse exemplo não eliminou por completo os subgrafos que rodeiam a estrutura central da Figura 5.1.1.

Nos exemplos anteriores, as cliques máximas dos grafos foram obtidas com auxílio da linguagem de programação *python*. Contudo, no cenário analisado, a ferramenta não foi capaz de identificar a clique máxima do grafo na Figura 5.1.2.1.

Como comentado anteriormente a identificação de cliques máximas em grafos é considerado um problema NP-difícil, o que significa que não existe nenhum algoritmo conhecido que garanta uma solução ótima para o problema em tempo polinomial. Devido a esse motivo, a ferramenta utilizada não foi capaz de identificar uma solução para o cenário analisado.

Logo, um novo filtro foi aplicado aos dados para possibilitar que as análises propostas fossem realizadas. O *dataset* pré-processado foi filtrado novamente e apenas as

relações com peso maior igual à 5 foram mantidas. O resultado desse processo pode ser visto na Figura 5.1.2.2.

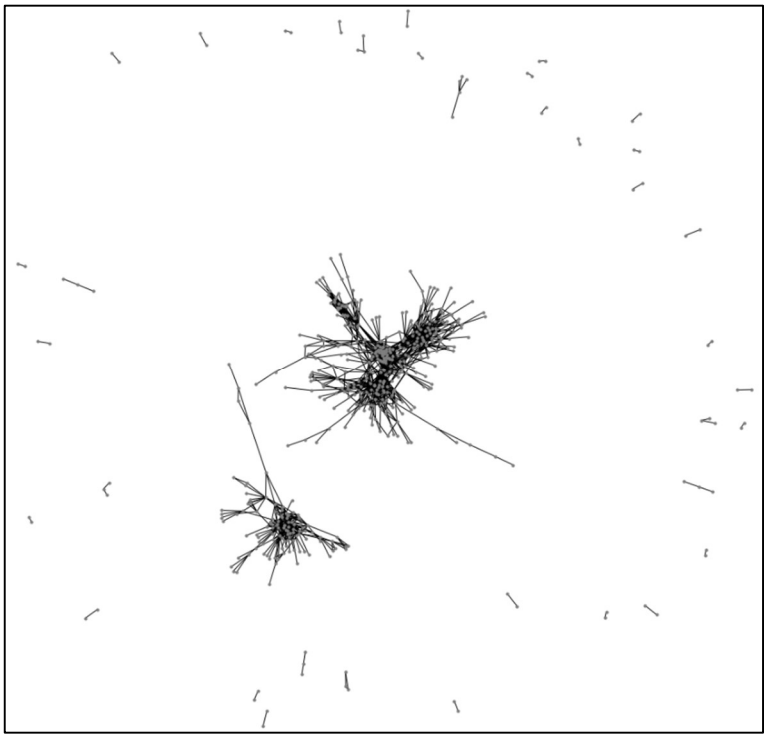


Figura 5.1.2.2 – Grafo das empresas com arestas com peso maior igual à 5.

O grafo reduzido e apresentado na Figura 5.1.2.2 possui 517 nós e 1.654 arestas. Ele é composto por três grandes agrupamentos: um formado por vários subgrafos pequenos que rodeiam as estruturas centrais e dois grandes subgrafos localizados no centro da Figura 5.1.2.2.

Para esse cenário de filtragem, o algoritmo utilizado conseguiu identificar uma clique máxima composta por 10 empresas. Ela é formada pelos nós 3480, 8849, 7364, 4036, 4277, 6742, 2916, 7049, 9709 e 8050 e é apresentada na Figura 5.1.2.3.

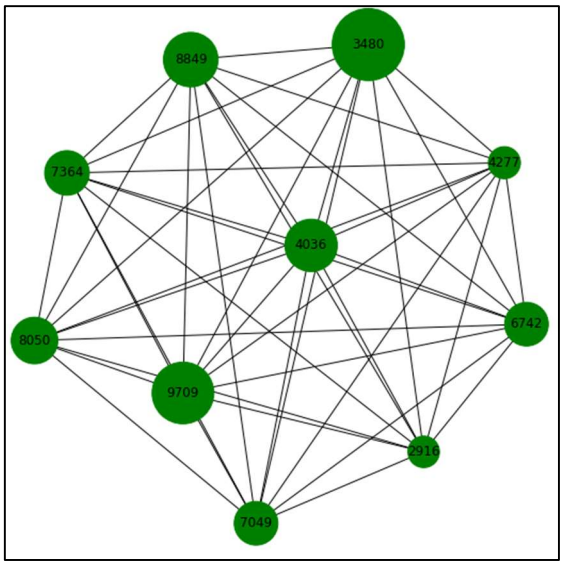


Figura 5.1.2.3 – Clique máxima do grafo da Figura 5.1.2.2.

Expandindo a clique máxima, foram adicionados 6 nós ao subgrafo para a geração da clique quase máxima, que representam as empresas 6136, 6302, 7121, 4648, 3180 e 9312. Essa estrutura apresenta uma densidade de 0,93 e pode ser visualizada na Figura 5.1.2.4.

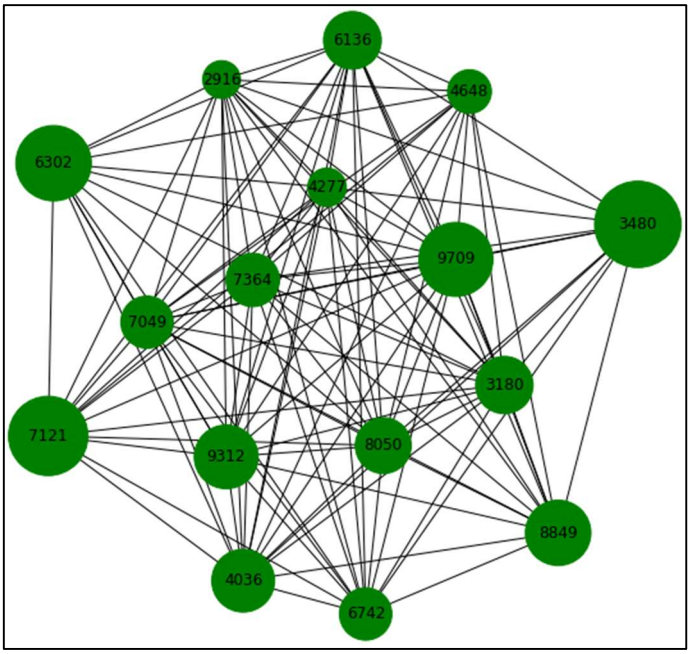


Figura 5.1.2.4 – Clique quase máxima do grafo da Figura 5.1.2.2.

Dentre os 6 nós adicionados à clique máxima para a geração do subgrafo mostrado na Figura 5.1.2.4, não houve a inclusão de nenhum nó vermelho.

Um último exemplo de aplicação do filtro proposto limitou as relações das empresas que permaneceriam no *dataset* para apenas aquelas com valor maior ou igual à 10. O resultado desse procedimento pode ser visualizado na Figura 5.1.2.5.

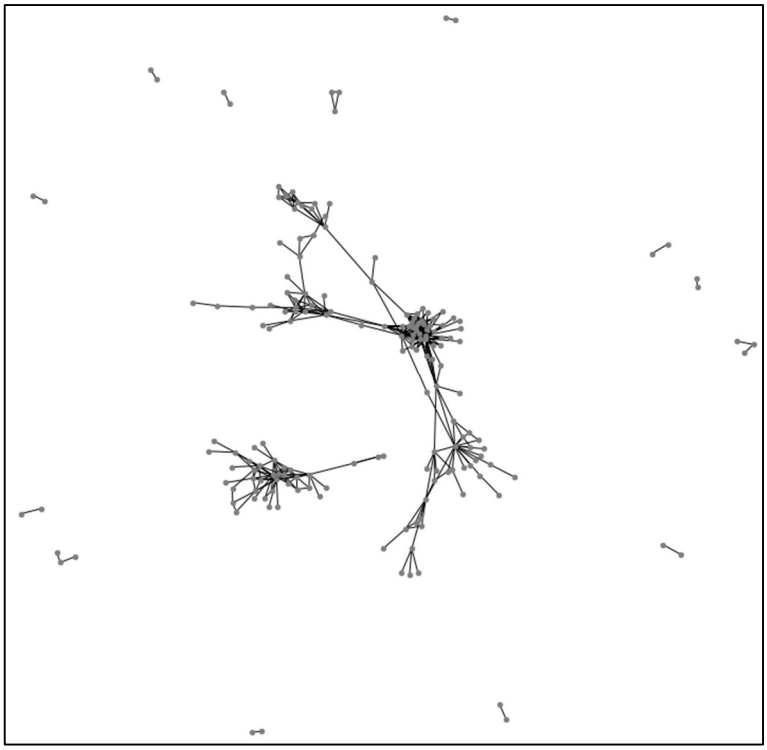


Figura 5.1.2.5 – Grafo das empresas com arestas com peso maior igual à 10.

Com a aplicação do novo cenário de filtragem, restaram 182 nós e 417 arestas no grafo. Apesar da grande redução realizada nos dados, nota-se ainda uma grande presença de pequenos subgrafos, além de dois grandes conjuntos de nós.

Aplicando a ferramenta para detecção de cliques máximas no grafo, foi identificado um subgrafo completo com 6 nós. As empresas que formam essa estrutura são as com código de identificação 7026, 6136, 7624, 4036, 9153 e 1315. A clique máxima pode ser visualizada na Figura 5.1.2.6.

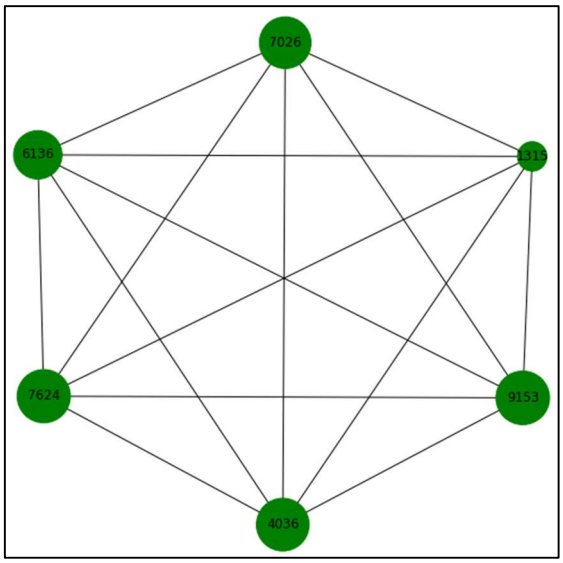


Figura 5.1.2.6 – Clique máxima do grafo da Figura 5.1.2.5.

A clique exibida na Figura 5.1.2.6 representa uma situação muito interessante. Cada aresta do subgrafo apresenta um peso maior ou igual à 10. Isso significa que cada uma das 6 empresas destacadas participou de pelo menos 10 licitações com cada uma das 5 outras companhias. Apesar de nenhuma das empresas terem sido alvo de investigação pelo TCE/PB, as relações destacadas são bastante curiosas e exigem uma atenção especial.

Para finalizar a análise desse exemplo de aplicação, a clique máxima mostrada na Figura 5.1.2.6 foi expandida a partir da adição de 6 nós, o que reduziu a densidade do subgrafo para 0,94. A estrutura obtida pode ser vista na Figura 5.1.2.7.

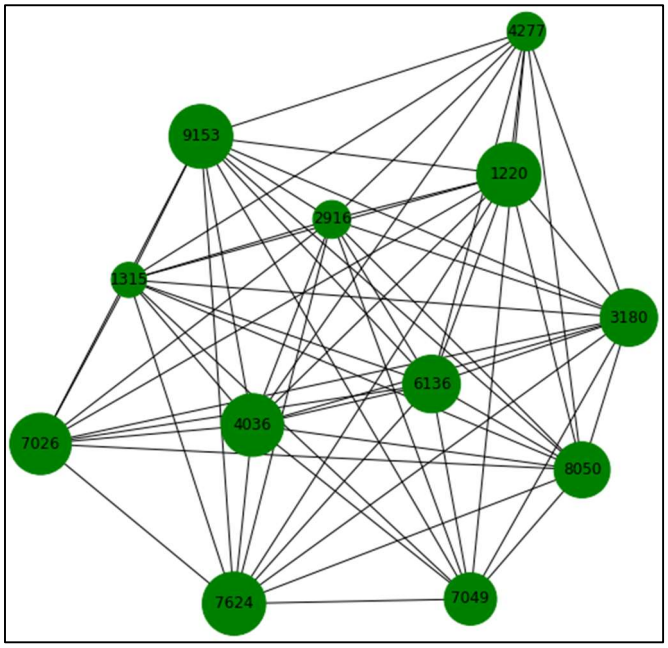


Figura 5.1.2.7 – Clique quase máxima do grafo da Figura 5.1.2.5.

As empresas adicionadas ao subgrafo da Figura 5.1.2.7 são identificadas pelos códigos 4277, 1220, 8050, 2916, 3180 e 7049. Analisando a clique quase máxima gerada, nota-se que mesmo 6 nós terem sido adicionados à clique máxima, a densidade do subgrafo ainda está bastante elevada, indicando que a estrutura gerada é bastante coesa.

Com base com resultados apresentados e nas análises realizadas, nota-se que filtro que busca limitar os dados com base na quantidade de relações mantidas entre as empresas é uma das inúmeras metodologias que podem ser empregadas para a redução do *dataset*.

5.1.3. Aplicação de filtro: analisando apenas empresas investigadas

Para finalizar a exemplificação de filtros que podem ser aplicados nos dados para facilitar a análise dos grafos, foram selecionados do *dataset* original apenas as informações relacionadas às empresas que já passaram por processos investigativos juntos ao TCE/PB. Com essas informações, foi possível gerar o grafo exibido na Figura 5.1.3.1.

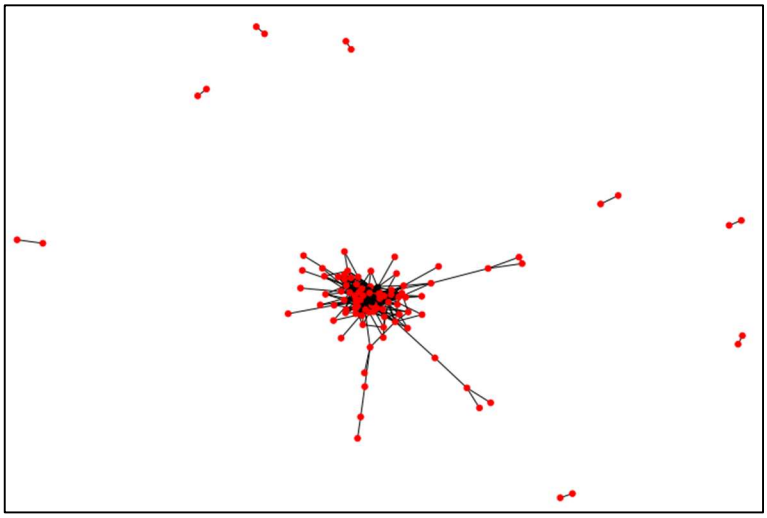


Figura 5.1.3.1 – Grafo gerado considerando apenas dados de empresas investigadas.

O grafo mostrado na Figura 5.1.3.1 é composto por 108 nós e por 492 arestas. Ele possui 9 subgrafos, sendo que o maior deles apresenta 92 nós. Esse subgrafo é destacado na Figura 5.1.3.2.

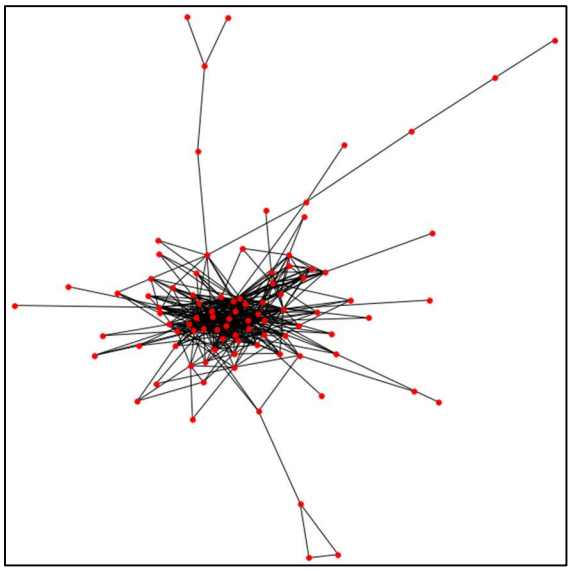


Figura 5.1.3.2 – Destaque para o maior subgrafo do grafo da Figura 5.1.3.2.

Apesar de em um primeiro momento o fato de o grafo da Figura 5.1.3.1 possuir apenas 108 nós parecer estranho, considerando que 202 empresas passaram por processos investigativos, essa informação pode ser facilmente compreendida.

No grafo são exibidas apenas as relações que empresas investigadas mantiveram entre si. Assim, caso uma empresa tenha sido investigada, mas durante o seu período de participações ela nunca tenha concorrido em uma mesma licitação com outra companhia que também tenha sido investigada, tal empresa não será exibida após a aplicação do filtro utilizado.

Prosseguindo com a análise, uma clique máxima foi extraída do grafo exibido na Figura 5.1.3.1. Esse subgrafo é mostrado na Figura 5.1.3.3.

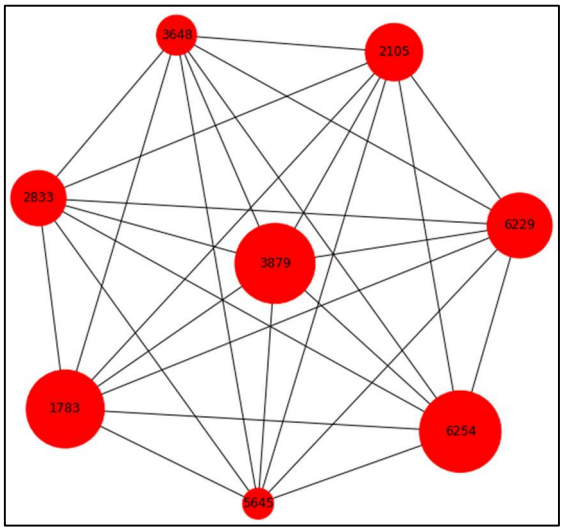


Figura 5.1.3.3 – Clique máxima do grafo da Figura 5.1.3.1.

A clique máxima obtida possui 8 nós e é constituída pelas empresas 3648, 2105, 2833, 3879, 6229, 1783, 5645 e 6254. Como foram utilizadas para a confecção do grafo da Figura 5.1.3.1 apenas dados de empresas investigadas, todos os nós da clique máxima foram sinalizados na cor vermelha.

Para a obtenção de uma clique quase máxima, foram adicionados 12 nós ao grafo mostrado na Figura 5.1.3.3. Para a realização da expansão da clique máxima, foram utilizadas informações do *dataset* completo que contém todas as empresas, investigadas ou não.

A opção por utilizar todas as informações para a obtenção de uma clique quase máxima buscou identificar empresas que não foram investigadas, mas que mantém grande relação com companhias que já passaram por processos investigados.

Logo, na Figura 5.1.3.4 a seguir está exibida a clique quase máxima obtida para o cenário de filtragem estudado. Foram adicionados 12 nós à clique máxima da Figura 5.1.3.3, gerando uma densidade de 0,95 para o grafo.

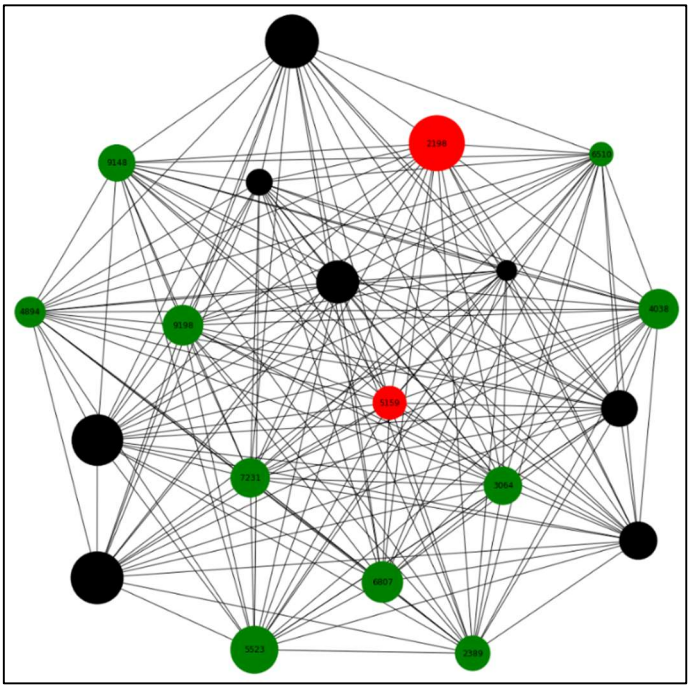


Figura 5.1.3.4 – Clique quase máxima do grafo da Figura 5.24.

Os nós adicionados à clique máxima mostrada na Figura 5.1.3.3 para a obtenção da clique quase máxima foram os das empresas 9148, 4894, 9198, 2198, 6510, 4038, 5159, 3064, 7231, 5523, 6807 e 2389.

Na clique quase máxima exibida na Figura 5.1.3.4, há nós pintados em três cores: em verde, que representam empresas que não foram investigadas pelo TCE/PB, em vermelho, que indicam companhias investigadas que foram adicionadas à clique máxima para a obtenção do novo subgrafo, e em preto, que compreendem as empresas investigadas originais à clique máxima.

Observando o resultado apresentado na Figura 5.1.3.4, nota-se que dentre os 12 nós adicionados para a obtenção da clique quase máxima, dois foram de empresas investigadas e 10 de empresas não investigadas. Assim, os nós não investigados adicionados para a formação da clique quase máxima podem ser assumidos como de alto potencial de investigação, devido às suas relações com empresas suspeitas.

5.1.4. Análise da estrutura dos grafos

Nos grafos apresentados anteriormente na seção 5.1.2, nota-se a formação de dois grandes grupos que são conectados por algumas ligações. Essa curiosa divisão pode ser facilmente compreendida quando as características das empresas que compõem cada conjunto são analisadas.

Como dito anteriormente, durante o processo de filtragem do *dataset*, as licitações foram classificadas de acordo com o seu objeto de contratação em 7 categorias: material/equipamento, abastecimento, pavimentação, obra, reforma, outros e projeto.

Assim, com o intuito de identificar a área de atuação principal das empresas, avaliou-se para cada entidade analisada qual o tipo de licitação que ela mais participava. Essa classificação forneceu a informação acerca do tipo principal da empresa.

Usando esse dado, os grafos mostrados na seção 5.1.2 foram plotados novamente, mas dessa vez os nós foram coloridos de acordo com a classificação da empresa. As cores atribuídas a cada uma das classes foram:

- Material/Equipamento → Vermelho
- Abastecimento → Amarelo
- Pavimentação → Laranja
- Obra → Azul
- Reforma → Roxo
- Outros → Rosa
- Projeto → Verde

Dessa forma, na Figura 5.1.4.1 a seguir são exibidos os grafos coloridos, com e sem arestas, segundo tal padrão de coloração e considerando a aplicação de filtros no *dataset* para a seleção dos nós que obtiveram mais de 10, 20 e 50 participações em licitações.

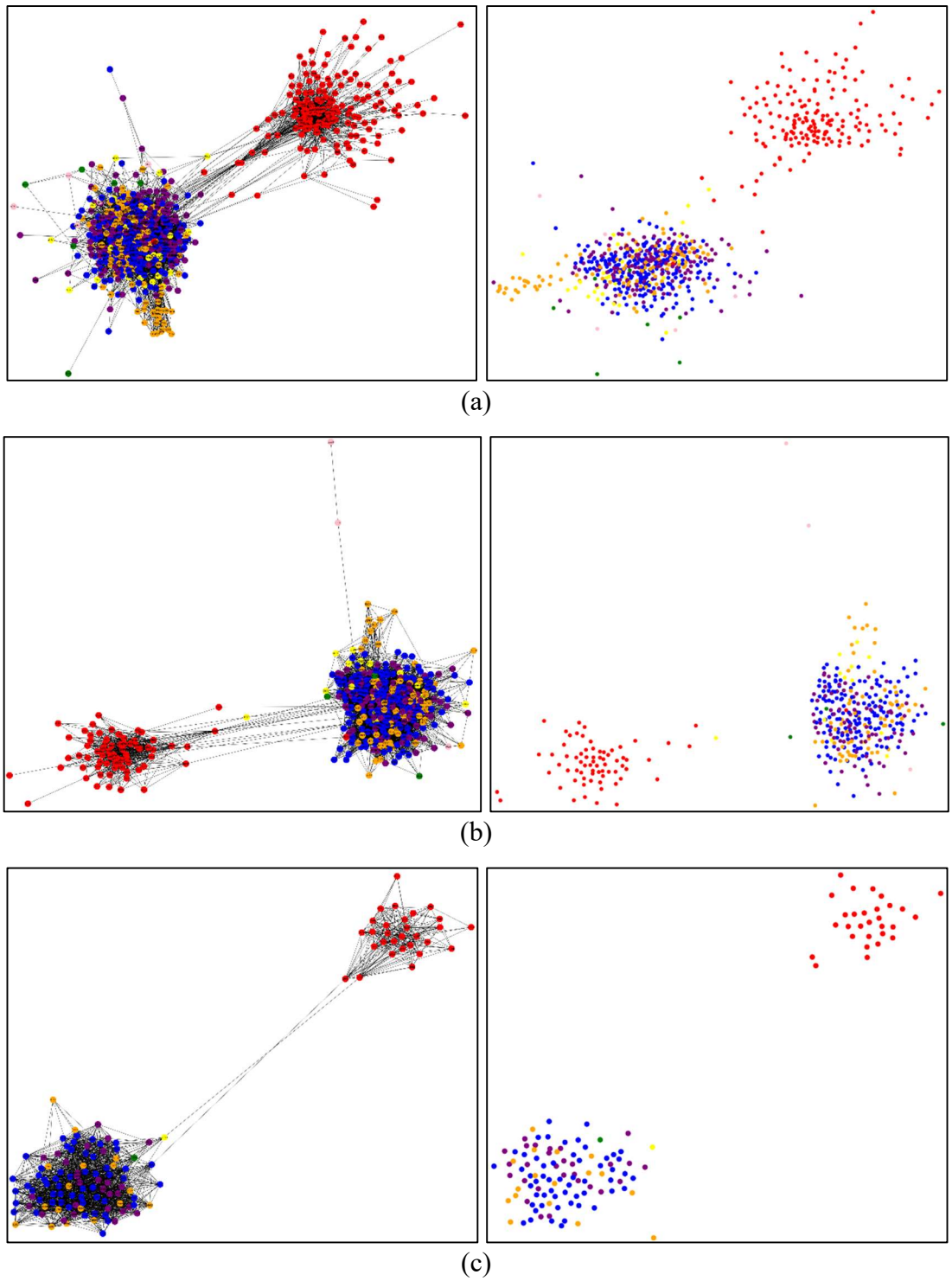


Figura 5.1.4.1 – Grafos coloridos de acordo com a classificação das empresas com e sem arestas considerando apenas os nós com mais de 10 (a), 20 (b) e 50 (c) participações em licitações.

Dessa forma, como mostrado na Figura 5.1.4.1, nota-se uma clara divisão no grafo gerada pelo tipo principal de atuação da empresa analisada. Enquanto de um lado do grafo há uma maior presença de empresas coloridas de vermelho, indicando uma área de atuação voltada para o fornecimento de materiais e de equipamentos, do outro lado do grafo há uma prevalência das empresas das demais áreas de atuação.

Assim, devido à essa clara característica de divisão dos dados e como o objetivo deste estudo é analisar padrões suspeitos que possam indicar a formação de conluio, análises futuras realizadas nesse mesmo *dataset* e com objetivo semelhante a esse podem ser realizadas separando as licitações de materiais e de equipamentos das demais relacionadas à construção civil.

5.2. CLUSTERIZAÇÃO

O segundo método aplicado neste trabalho consistiu no agrupamento das empresas que participaram de licitações com base nas suas características de atuação nos certames. Foi utilizada a técnica de clusterização k-means em conjunto com o PSO, assim como apresentado na seção 3.2.

Para isso, o k-means foi aplicado durante 300 iterações ou até que os resultados obtidos não variassem mais do que 0,0001 unidades entre as repetições. Já para a execução do PSO, foram adotadas 20 partículas e foram executadas 150 iterações. O valor de *w*, o peso de inércia, sofreu um decaimento linear, de modo que o valor da velocidade da partícula de uma iteração corresponde à 98% da iteração anterior.

Para a realização da clusterização, foi utilizado o *dataset* reduzido apresentado na seção 4.2. Assim, foram selecionadas apenas as informações dos certames da modalidade “Tomada de Preços”, cujo tipo era “Obra” ou “Reforma”, que tinham mais de 1 e menos de 8 concorrentes e que não eram licitações de lote.

A redução realizada no *dataset* pré-processado e utilizado na seção anterior foi necessária tendo em vista que esse conjunto de dados apresentava algumas inconsistências relacionadas aos custos das licitações. Como esses parâmetros serão de grande importância para o agrupamento dos dados, foi preciso que uma correção manual das informações fosse realizada.

Visto que o *dataset* original apresentava mais de 30 mil informações de 14 mil certames, a correção dos preços de todo o conjunto de dados era inviável. Por isso, optou-se

por reduzir o *dataset* original. Apesar disso, essa diminuição não irá prejudicar os resultados que serão apresentados neste trabalho.

O *dataset* reduzido apresenta 10 informações sobre os dados, que são: o número de participações da empresa, sua taxa de vitórias, o total proposto em licitações, o total vencido, a relação percentual entre a quantia vencida e a proposta, a quantidade de anos que a empresa participou de certames, o número médio de licitações concorridas por ano, o total de cidades que a empresa concorreu e a quantidade de municípios que ela venceu pelo menos uma licitação (vide Tabela 4.1.2).

No entanto, a utilização dessas 10 variáveis no processo de clusterização dos dados é inviável. Isso ocorre, pois, em um *dataset* com muitos parâmetros, há uma grande chance de uma ou mais variáveis estarem relacionadas a uma mesma informação. Assim, utilizar ambos os parâmetros acarretaria em um aumento do custo computacional exigido sem garantia que uma melhora significativa no resultado ocorresse.

Além disso, em um *dataset* com muitas variáveis, um parâmetro que seja mais relevante ao conjunto de dados pode perder sua importância caso muitas variáveis irrelevantes ou menos significativas sejam analisadas. Ademais, como serão usados vários cenários de manipulação das variáveis para avaliar os diferentes comportamentos dos dados, caso todas as 10 variáveis fossem utilizadas, seria necessária a realização de mais de 1000 agrupamentos do *dataset*.

Logo, devido aos fatores supramencionados optou-se por reduzir o conjunto de variáveis analisadas de 10 para apenas 4 parâmetros. Dessa forma, a quantidade de cenários de agrupamento que deveriam ser examinados diminuiu de aproximadamente 1000 para somente 11.

Para realizar a seleção das variáveis que permaneceriam no estudo, foi utilizada a técnica de redução de dimensionalidade PCA. Destaca-se que o uso desse método para diminuir a quantidade de variáveis analisadas não está relacionado com a utilização do PCA antes da aplicação do k-means no processo de clusterização, conforme apresentado na seção 3.2.2.

Para selecionar as 4 variáveis que seriam utilizadas para a clusterização dos dados, foram construídos, utilizando as variáveis originais, 10 subgrupos compostos por 9 variáveis cada. Em cada um desses conjuntos uma variável do grupo original foi excluída.

Em seguida, em cada um dos 10 subgrupos foi aplicado PCA para reduzir as 9 variáveis para apenas duas. Após a aplicação do PCA, para cada cenário testado, foi

calculada a precisão da redução dos dados, ou seja, quanto as duas novas variáveis geradas pelo PCA representavam do conjunto original.

Após isso, o valor das precisões calculadas foi comparado e o conjunto com o maior valor foi selecionado. A variável que foi excluída do conjunto original para a construção desse grupo foi removida da análise. Esse procedimento foi aplicado novamente, considerando o grupo obtido a partir da exclusão da variável como o novo conjunto original, até que restassem apenas 4 variáveis.

Para uma melhor compreensão do procedimento adotado para a seleção das 4 variáveis que serão utilizadas na clusterização do *dataset* reduzido, é apresentado na Figura 5.2.1 um esquema resumindo os passos descritos.

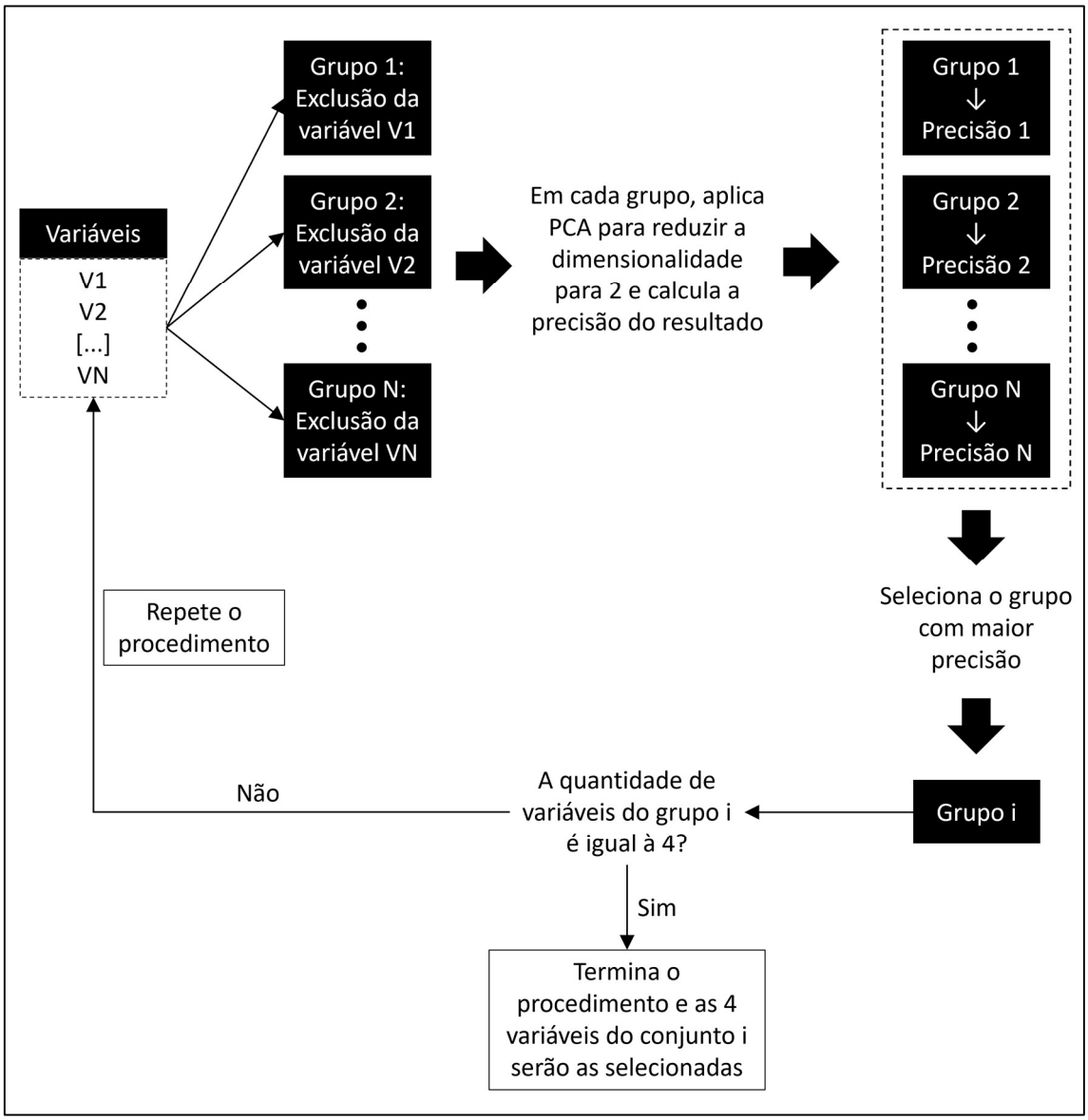


Figura 5.2.1 – Esquema do procedimento utilizado na redução das variáveis do *dataset*.

Logo, após a aplicação do procedimento descrito, restaram no *dataset* quatro variáveis, que foram o número de vitórias da empresa, sua taxa de vitórias, o total proposto em licitações e a quantidade de cidades que a companhia venceu certames. Usando essas informações, foram elaborados 11 cenários de agrupamento, que podem ser visualizados na Tabela 5.2.1.

Tabela 5.2.1 – Cenários de agrupamento que serão utilizados.

Cenário	N. vitórias	Taxa de vitórias	Total proposto	N. Cidades vitoriosas
1	X	X	X	X
2	X	X	X	
3	X	X		X
4	X		X	X
5		X	X	X
6	X	X		
7	X		X	
8	X			X
9		X	X	
10		X		X
11			X	X

Os cenários de agrupamento mostrados na Tabela 5.2.1 foram utilizados para a aplicação da clusterização utilizando o k-means em conjunto com o PSO.

Logo, utilizando o *dataset* reduzido, as variáveis selecionadas, os cenários de agrupamento gerados e o procedimento apresentado na seção 3.2.2, foi possível realizar a clusterização do conjunto de dados utilizando os algoritmos k-means e PSO em conjunto.

Para a aplicação do k-means, é necessário que a quantidade de clusters seja determinada inicialmente. Neste trabalho, foi utilizado o método de Elbow para a realização dessa determinação. Aplicando o método em cada um dos cenários de agrupamento propostos, verificou-se que em alguns casos o conjunto de dados tendia para a criação de 3 clusters, enquanto em outros havia uma maior tendência à formação de 4 grupos. Contudo, como almeja-se comparar os resultados apresentados por cada um dos cenários de agrupamento, seria necessário que em todos os casos fosse adotado um mesmo valor inicial para o número de clusters.

Logo, como parte dos cenários de agrupamento tendia para a formação de 3 clusters e parte para 4 grupos, optou-se por realizar o processo de clusterização duas vezes em cada situação analisada. A seguir, serão mostrados os resultados obtidos para cada um dos cenários de agrupamento adotados para cada quantidade de clusters utilizada.

5.2.1. RESULTADOS PARA 3 CLUSTERS

Nesta seção serão apresentados os resultados das clusterizações realizadas adotando 3 clusters. Logo, na Figura 5.2.1.1 a seguir são mostrados os agrupamentos dos dados para o primeiro, o segundo, o terceiro e o quarto cenários analisados.

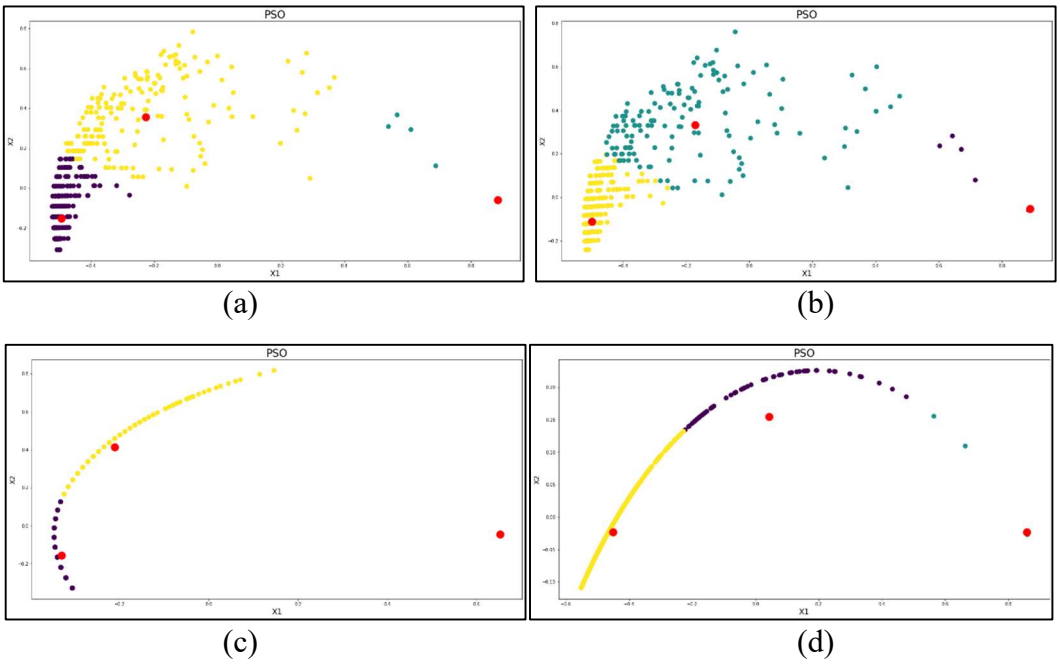


Figura 5.2.1.1 – Resultados da clusterização dos cenários 1 (a), 2 (b), 3 (c) e 4 (d) usando 3 clusters.

Os agrupamentos mostrados na Figura 5.2.1.1 apresentaram, após a otimização dos centroides utilizando o PSO, erros de quantização de 2.25 (cenário 1), 2.08 (cenário 2), 1.74 (cenário 3) e 1.32 (cenário 4).

Realizando a clusterização dos dados considerando os cenários de agrupamento 5, 6, 7 e 8, foram obtidos os resultados expostos na Figura 5.2.1.2.

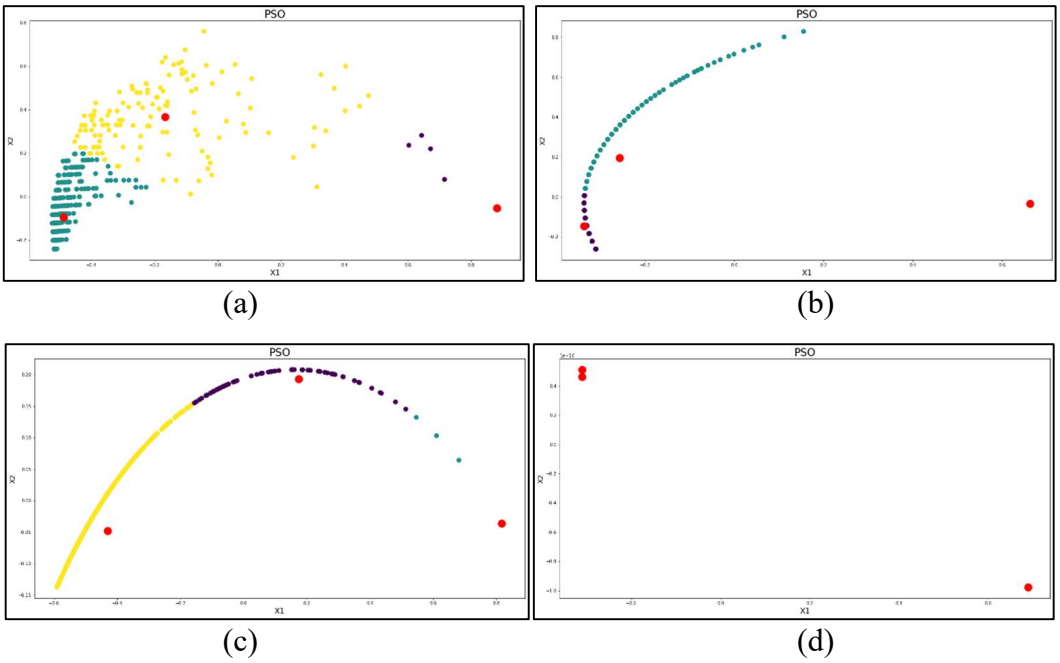


Figura 5.2.1.2 – Resultados da clusterização dos cenários 5 (a), 6 (b), 7 (c) e 8 (d) usando 3 clusters.

Para os cenários 5, 6, 7 e 8, os erros de quantização obtidos após a aplicação do algoritmo de clusterização foram de 2.09, 1.65, 1.62 e 0, respectivamente.

Por fim, a Figura 5.2.1.3 apresenta os resultados da aplicação do k-means combinado com o PSO no *dataset* reduzido para os cenários de agrupamento 9, 10 e 11.

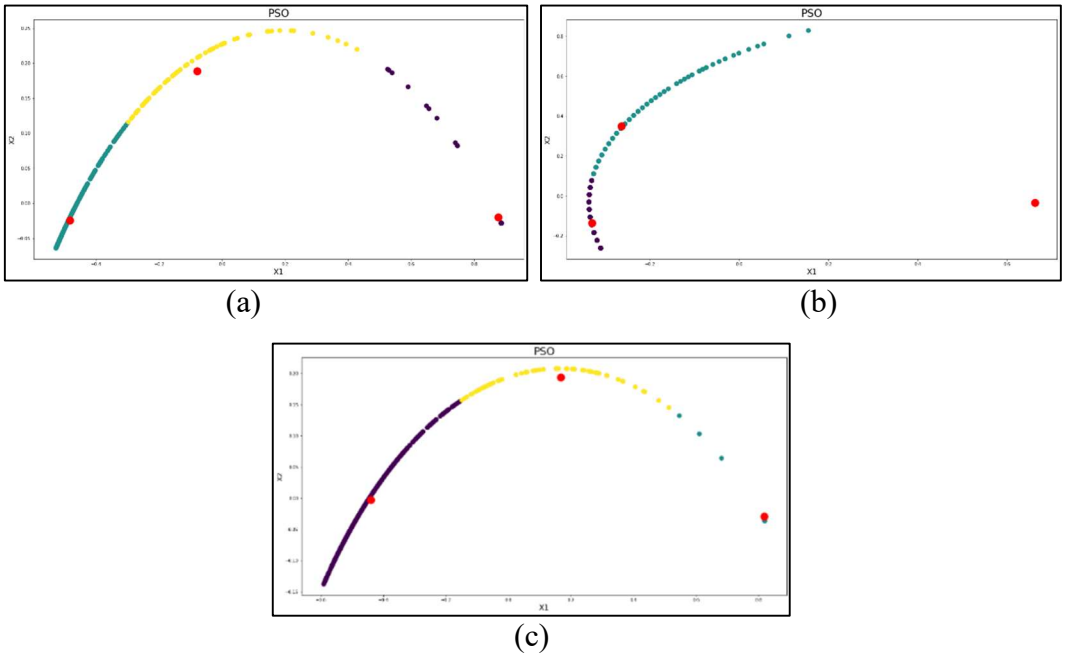


Figura 5.2.1.3 – Resultados da clusterização dos cenários 9 (a), 10 (b) e 11 (c) usando 3 clusters.

Os cenários 9, 10 e 11 apresentaram erros de quantização de 1.33, 1.53 e 1.56, de forma respectiva, após a clusterização utilizando k-means combinado com PSO.

Analisando os resultados da clusterização para os 11 cenários de agrupamento exibidos nas Figuras 5.2.1.1, 5.2.1.2 e 5.2.1.3, nota-se que alguns dos gráficos aparentam possuir menos pontos do que empresas analisadas. Assim, é importante ressaltar que algumas companhias, a depender do seu comportamento de atuação nas licitações e das variáveis consideradas na análise, apresentam as mesmas coordenadas x e y geradas pela normalização e pela aplicação do PCA nos dados. Dessa forma, mais de uma empresa pode ser simbolizada por um mesmo ponto.

Após a finalização da aplicação do k-means em conjunto com o PSO nos 11 cenários de agrupamentos elaborados, cada um dos casos analisados dividiu os dados em três grupos. Porém, essa divisão não foi realizada exatamente da mesma forma em todos os cenários.

Isso ocorre, pois, ao considerar-se determinados parâmetros, uma empresa pode ser alocada em um grupo, mas ao examinar-se outras variáveis tal companhia acaba sendo classificada em outro conjunto. Além disso, como o algoritmo identifica os grupos de modo aleatório, o primeiro cluster de um cenário não corresponde, necessariamente, ao primeiro agrupamento dos demais conjuntos.

Logo, após a clusterização dos dados, foram gerados 11 resultados, a priori, diferentes. Como o objetivo final desse procedimento de agrupamento dos dados é separar as empresas em grupos com base em seus padrões de comportamento, foi preciso unificar os resultados obtidos pelos cenários de clusterização analisados.

Para isso, primeiro foi necessário padronizar a nomenclatura dos grupos entre os cenários estudados. Ao realizar a divisão das empresas nos conjuntos, o algoritmo analisado nomeou os grupos de modo aleatório de “0”, “1” e “2”. Assim, antes de realizar a unificação dos resultados, foi preciso identificar as relações entre as classificações. Para uma melhor compreensão do procedimento realizado, serão utilizados como exemplo os resultados da clusterização para os cenários de agrupamento 1 e 2, exibidos na Tabela 5.2.1.1.

Tabela 5.2.1.1 – Resultados da clusterização para os cenários de agrupamento 1 e 2, considerando 3 clusters.

Cenário/Classificação	0	1	2
1	327	230	171
2	230	142	356

Analisando as informações da Tabela 5.2.1.1 e comparando as empresas de cada um dos grupos criados para cada um dos cenários analisados, foi verificado que 327 das 356 companhias do grupo “2” do segundo cenário estavam localizadas na classe “0” do primeiro cenário.

Prosseguindo com o estudo, constatou-se que as 230 empresas alocadas na categoria “1” do cenário 1 eram as mesmas 230 companhias do grupo “0” do cenário de agrupamento 2. Já as 171 empresas do grupo “2” do cenário 1 correspondiam a 142 companhias do grupo “1” e a 29 da classe “2”, ambos do cenário 2.

Verificou-se, portanto, que os grupos “0”, “1” e “2” do primeiro cenário de agrupamento correspondiam às categorias “2”, “0” e “1” do segundo cenário, de forma respectiva. Essa análise foi realizada também para os demais cenários de agrupamento estudados, de modo a padronizar a nomenclatura das classes entre os conjuntos.

Em seguida, para cada uma das 728 empresas analisadas, foi avaliado as diferentes classificações atribuídas às companhias ao longo dos 11 cenários de agrupamento adotados afim de determinar a qual grupo elas pertenciam.

Como comentado, uma empresa pode ser classificada em diferentes grupos em cenários de clusterização diferentes, a depender das variáveis consideradas em cada caso. Para determinar qual conjunto melhor correspondia ao padrão de comportamento de uma companhia, optou-se por alocar as empresas no grupo em que elas foram mais classificadas ao longo dos 11 cenários analisados.

Por exemplo, caso uma determinada empresa tenha sido classificada no grupo 1 em 9 cenários de agrupamento, mas tenha sido inclusa 2 vezes no conjunto 0, a companhia exemplificada será alocada no grupo 1, pois foi a classificação que prevaleceu.

Dessa forma, a partir da clusterização dos dados usando k-means e PSO e adotando três clusters, as 728 empresas do *dataset* reduzido foram divididas em três grupos, denominados A, B e C, com base em seu padrão de comportamento nas licitações examinadas.

Assim, na Tabela 5.2.1.2 são apresentadas as principais características do primeiro conjunto identificado, o grupo A.

Tabela 5.2.1.2 – Grupo A formado pela clusterização dos dados usando 3 grupos.

Variável	Mínimo	Máximo	Médio
Nº de participações	1	13	4,66

Variável	Mínimo	Máximo	Médio
Nº de vitórias	1	9	2,22
Taxa de vitórias	11,11%	100%	59,72%
Total proposto	-	-	R$ 358.904,19 / licitação
Total vencido	-	-	R$ 367.032,24 / licitação
Qtd de anos	1	6	2,12
Nº de licitações por ano	1	10	2,23
Nº de cidades	1	10	2,95

No grupo A há 362 empresas, sendo que 31 delas já foram investigadas. Analisando esses valores, verifica-se que a taxa de investigação do grupo A, que é uma relação entre a quantidade de companhias investigadas e o total de empresas que compõem do grupo, vale 8,56%.

As empresas do grupo A, de acordo com as informações apresentadas na Tabela 5.2.1.2, apresentam baixo número médio de participações, especialmente quando comparado com os outros grupos, mas uma elevada taxa média de vitórias, que é a maior de todos os conjuntos.

Analisando a quantidade média de anos em que as empresas do grupo A participaram de licitações, verifica-se um valor relativamente baixo, considerando que o *dataset* estudado contempla 8 anos de dados de licitações. Essas companhias também apresentaram um baixo valor médio de participações em licitações por ano e de cidades onde concorreram em certames.

Na Tabela 5.2.1.3 a seguir são apresentadas as características do segundo grupo de empresas elaborado usando os resultados da aplicação do k-means com o PSO.

Tabela 5.2.1.3 – Grupo B formado pela clusterização dos dados usando 3 grupos.

Variável	Mínimo	Máximo	Médio
Nº de participações	1	46	2,70
Nº de vitórias	0	4	0,02
Taxa de vitórias	0,00%	8,70%	0,06%
Total proposto	-	-	R$ 676.717,57 / licitação
Total vencido	-	-	R$ 902.194,17 / licitação
Qtd de anos	1	7	1,41

Variável	Mínimo	Máximo	Médio
Nº de licitações por ano	1	11,50	1,66
Nº de cidades	1	16	1,80

Participam do grupo B 228 empresas, sendo que apenas 7 delas já foram investigadas pelo TCE/PB, o que representa uma taxa de investigação de apenas 3,07%.

Observando a Tabela 5.2.1.3 e analisando o número médio de participações das empresas, seu número médio de vitórias e sua taxa média de vitórias, constata-se que as companhias do grupo B apresentam baixa taxa de participação em licitações e quase nunca vencem os certames que participam.

Chamam atenção as altas quantias do valor proposto em média por licitação e do valor vencido em média por certame, que são quase o dobro dos demais conjuntos. Isso mostra que as empresas do grupo B quase nunca vencem licitações, mas quando ganham são em certames de preço bastante elevado.

Como já seria esperado devido ao baixo número de participações, as empresas do grupo B concorrem em média durante poucos anos, apresentando uma baixa taxa anual de participações em licitações. Essas companhias também possuem a característica de concorrerem em poucas cidades.

Por fim, na Tabela 5.2.1.4 são apresentadas as principais características das empresas que foram alocadas no grupo C.

Tabela 5.2.1.4 – Grupo C formado pela clusterização dos dados usando 3 grupos.

Variável	Mínimo	Máximo	Médio
Nº de participações	7	93	23,02
Nº de vitórias	1	25	5,69
Taxa de vitórias	5,45%	72,22%	25,43%
Total proposto	-	-	R$ 431.090,48 / licitação
Total vencido	-	-	R$ 391.831,38 / licitação
Qtd de anos	2	8	4,86
Nº de licitações por ano	1,75	18	4,84
Nº de cidades	1	55	11,85

O grupo C é composto por 138 empresas, sendo que 28 delas já foram alvo de processo investigativo movido pelo TCE/PB, representando uma taxa de investigação de 20,29%, o maior valor dentre os três grupos.

Analisando a Tabela 5.2.1.4, nota-se que as empresas que participam do grupo C apresentam alto número de participações, mas uma taxa de vitórias baixa. Com relação aos valores praticados nas licitações, verifica-se que eles são um pouco maiores do que os do grupo A, mas bem menores que os do B.

Dentre os três grupos identificados, as empresas do grupo C foram as que participaram de licitações durante o maior período de tempo e também as que apresentaram a maior taxa de participação em certames por ano. Essas companhias também concorreram em licitações de um maior número de cidades do que as empresas dos outros grupos.

Logo, de forma resumida, os grupos que dividem as empresas do *dataset* reduzido com base em seu padrão de comportamento e que foram obtidos com a aplicação do k-means em conjunto com o PSO considerando a formação de 3 clusters podem ser caracterizados da seguinte forma:

- Grupo A: compreende empresas com baixas taxas de participação e com altas taxas de vitórias;
- Grupo B: estão inclusas empresas com baixas taxas de participação e com taxa de vitórias praticamente nula;
- Grupo C: composto por empresas com altas taxas de participação em licitações e com baixas taxas de vitórias.

Analisando as características dos três grupos de empresas criados, verifica-se a formação de dois conjuntos que podem ser considerados como mais suspeitos de cometerem fraudes, que são os grupos A e C.

As empresas do grupo A apresentam uma taxa de vitórias maior do que seria esperado em um cenário de concorrência real. Essa situação, associada a uma baixa taxa de participações em licitações, levanta dúvidas acerca da legitimidade desse comportamento. O grupo C também apresenta um comportamento estranho, visto que ele é composto por empresas que participam de muitos processos licitatórios e que mesmo não apresentando uma boa taxa de vitórias, continuam competindo em certames.

É importante destacar que caso uma empresa faça parte de um grupo ou de outro, não significa que ela tenha cometido irregularidades. Entretanto, dúvidas quanto à sua

idoneidade podem surgir, especialmente quando associado a outros indícios de ocorrência de crime.

Salienta-se ainda que apesar de o grupo B não ser apontado como suspeito, ele não está isento da ocorrência de irregularidades. A classificação das empresas em grupos suspeitos é apenas um dos fatores que podem ser considerados na determinação da real suspeição de uma companhia. Assim, uma empresa pode ser culpada de cometer fraudes em licitações mesmo tendo sido classificada no grupo B.

5.2.2. RESULTADOS PARA 4 CLUSTERS

Nesta seção, será realizando novamente o procedimento apresentado na seção anterior, mas considerando dessa vez a divisão dos dados das empresas em 4 clusters. Destaca-se que a disposição dos dados em cada um dos cenários estudados é a mesma independente de se considerar a formação de 3 ou 4 clusters. A quantidade de grupos considerados irá influenciar apenas na divisão das empresas nos conjuntos elaborados.

Assim, na Figura 5.2.2.1 são exibidos os agrupamentos dos dados para o primeiro, o segundo, o terceiro e o quarto cenários analisados.

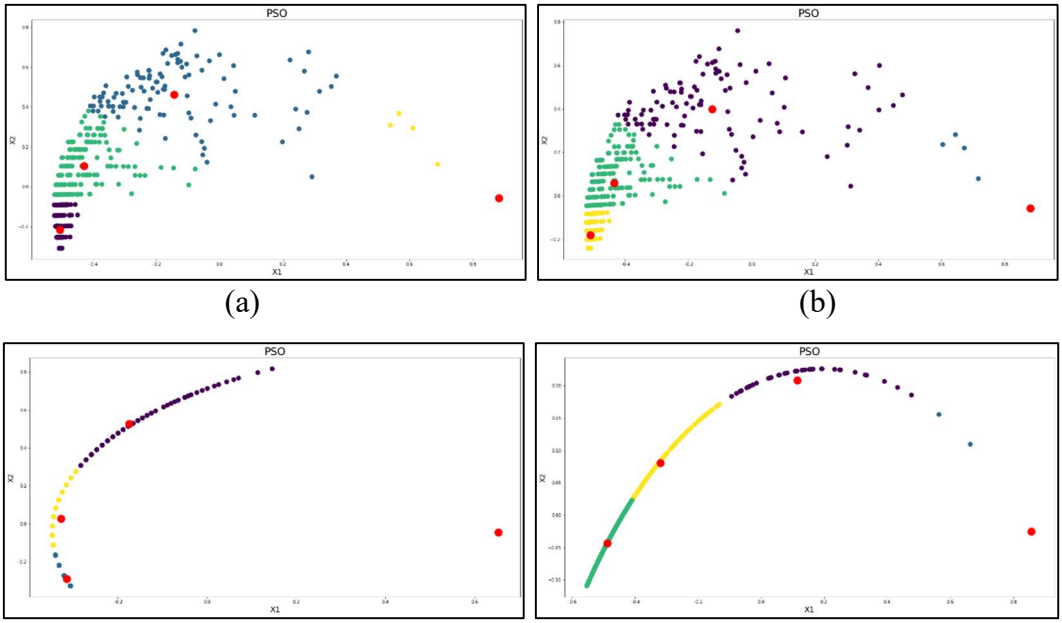
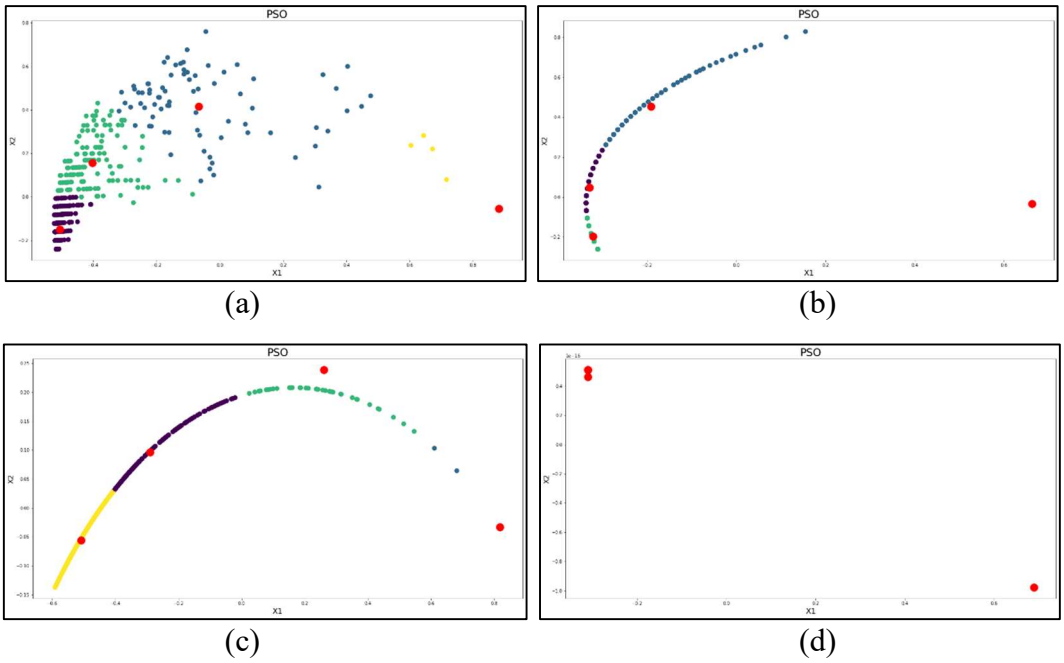


Figura 5.2.2.1 – Resultados da clusterização dos cenários 1 (a), 2 (b), 3 (c) e 4 (d) usando 4 clusters.

Os cenários de agrupamento 1, 2, 3 e 4 expostos na Figura 5.2.2.1 apresentaram erro de quantização de 1.55, 1.44, 1.10 e 0.83, de forma respectiva. A Figura 5.2.2.2 exibe os resultados da clusterização do k-means combinado com o PSO para os cenários de agrupamento 5, 6, 7 e 8.



(a) (b)

(c) (d)

Figura 5.2.2.2 – Resultados da clusterização dos cenários 5 (a), 6 (b), 7 (c) e 8 (d) usando 4 clusters.

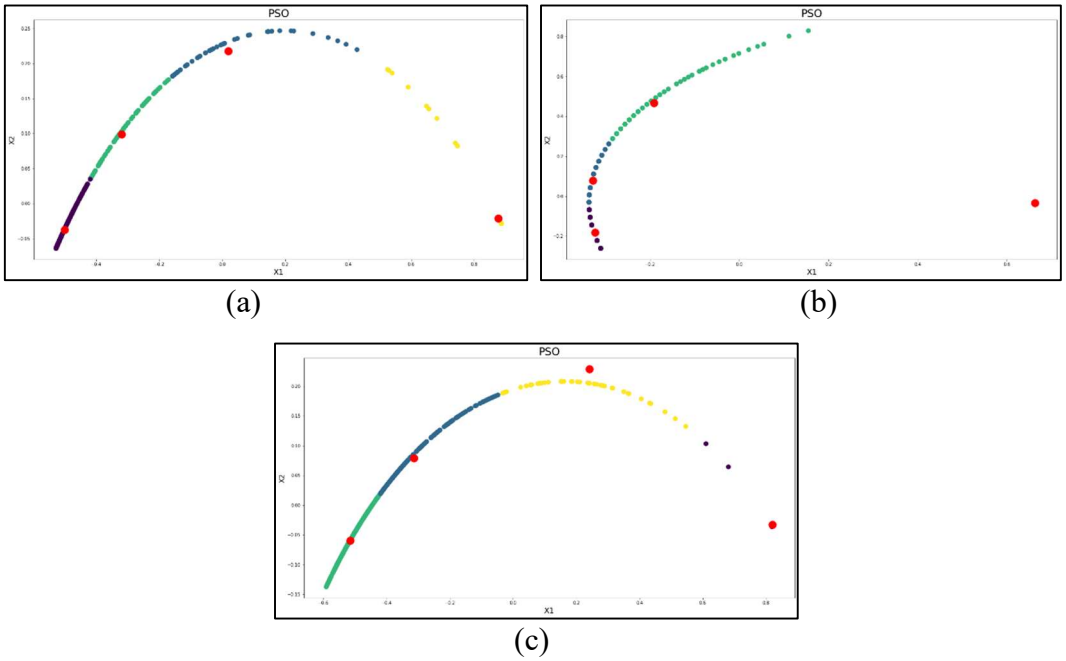
Os erros de quantização obtidos pelos cenários 5, 6, 7 e 8 após a aplicação do algoritmo de clusterização foram de 1.45, 0.94, 0.97 e 0, respectivamente.

O oitavo cenário de agrupamento apresentou um resultado diferente dos demais para o processo de clusterização usando 4 clusters. A aplicação do k-means em conjunto com o PSO identificou, como seria esperado, quatro centroides para representar cada um dos grupos gerados. Contudo, dos quatro centroides elaborados, dois apontam para o mesmo ponto. Ou seja, mesmo tendo sido solicitado ao algoritmo que ele dividisse as empresas em quatro grupos, devido à disposição espacial dos dados, a ferramenta conseguiu identificar apenas três conjuntos.

Analisando a disposição dos pontos mostrada na Figura 5.2.2.2 (d) e considerando que o erro de quantização obtido pelo algoritmo foi igual a 0, verifica-se que, ao analisar-se as variáveis “N. vitórias” e “N. de cidades vitoriosas”, a quantidade ideal de grupos é de fato três. Apesar de para esse cenário de agrupamento o quarto grupo de dados ser um conjunto

vazio, o resultado final obtido com a clusterização não será afetado, devido às análises realizadas após a execução de todos os agrupamentos.

Para finalizar a aplicação do k-means combinado com o PSO no *dataset* reduzido considerando a formação de 4 clusters, na Figura 5.2.2.3 são apresentados os resultados da clusterização para os cenários de agrupamento 9, 10 e 11.



(a) (b)

(c)

Figura 5.2.2.3 – Resultados da clusterização dos cenários 9 (a), 10 (b) e 11 (c) usando 4 clusters.

Foram obtidos erros de quantização de 0.90 para o cenário 9, de 0.95 para o cenário 10 e de 0.97 para o cenário de agrupamento 11.

Após a clusterização do *dataset* reduzido considerando os cenários de agrupamento elaborados e utilizando 4 clusters, foi possível aplicar o procedimento já descrito para a padronização das categorias dos grupos e para a unificação das classificações das empresas. Dessa forma, foram gerados quatro grupos de classificação das empresas, denominados de conjuntos A, B, C e D. Na Tabela 5.2.2.1 a seguir são apresentadas as principais características das companhias alocadas no grupo A.

Tabela 5.2.2.1 – Grupo A formado pela clusterização dos dados usando 4 grupos.

Variável	Mínimo	Máximo	Médio
Nº de participações	1	21	3,44

Variável	Mínimo	Máximo	Médio
Nº de vitórias	1	13	1,93
Taxa de vitórias	16,67%	100%	67,94%
Total proposto	-	-	R$ 320.204,43 / licitação
Total vencido	-	-	R$ 333.163,02 / licitação
Qtd de anos	1	6	1,78
Nº de licitações por ano	1	6	1,89
Nº de cidades	1	13	2,27

Participam do grupo A 263 empresas, sendo que 19 delas já foram alvo de processo investigativo pelo TCE/PB. Essa relação gera uma taxa de investigação de 7,22%.

Analisando as informações da Tabela 5.2.2.1, verifica-se que as empresas do grupo A apresentam um número de participações e um total de vitórias baixo, mas uma elevada taxa de vitórias. Esse conjunto também apresentou as características de participar de licitações durante um curto período de tempo e de concorrer em poucos certames por ano e em um número limitado de cidades.

As características do grupo B estão exibidas na Tabela 5.2.2.2.

Tabela 5.2.2.2 – Grupo B formado pela clusterização dos dados usando 4 grupos.

Variável	Mínimo	Máximo	Médio
Nº de participações	1	25	2,51
Nº de vitórias	0	1	0,00
Taxa de vitórias	0,00%	4,17%	0,02%
Total proposto	-	-	R$ 647.016,61 / licitação
Total vencido	-	-	R$ 568.623,68 / licitação
Qtd de anos	1	7	1,39
Nº de licitações por ano	1	11,50	1,64
Nº de cidades	1	16	1,80

O grupo B é constituído por 227 empresas e apresenta uma taxa de investigação de 3,08%, contendo apenas 7 companhias investigadas pelo TCE/PB.

A partir dos dados da Tabela 5.2.2.2, é possível compreender que o grupo B representa empresas com baixo número de participações em licitações e com uma taxa de

vitórias praticamente nula. Essas companhias também tiveram uma atuação temporal e espacial bastante limitada.

Examinando as informações relativas ao grupo B, nota-se que os valores propostos e vencidos médios por licitação são bastante elevados, especialmente quando comparado com os dos demais conjuntos de empresas.

A Tabela 5.2.2.3 traz as principais informações das empresas do grupo C.

Tabela 5.2.2.3 – Grupo C formado pela clusterização dos dados usando 4 grupos.

Variável	Mínimo	Máximo	Médio
Nº de participações	9	93	27,36
Nº de vitórias	1	25	6,19
Taxa de vitórias	5,45%	65,22%	23,13%
Total proposto	-	-	R$ 450.901,40 / licitação
Total vencido	-	-	R$ 410.113,16 / licitação
Qtd de anos	2	8	5,33
Nº de licitações por ano	1,8	18	5,31
Nº de cidades	1	55	13,70

O grupo C é formado por 97 empresas. Desse total, 23 foram investigadas, o que gera uma taxa de investigação de 23,71%, maior valor entre os quatro conjuntos analisados. O grupo C também apresenta o maior número médio de participações e o maior número de cidades onde as empresas concorreram em licitações.

Apesar do alto número de participações, as empresas do grupo C possuem uma taxa de vitórias bastante baixa. Essas companhias possuem altos valores para a quantidade média de anos de participação em licitações no estado da Paraíba e de número médio de concorrências em certames por ano.

Por último, na Tabela 5.2.2.4 a seguir são exibidas as principais informações do grupo D.

Tabela 5.2.2.4 – Grupo D formado pela clusterização dos dados usando 4 grupos.

Variável	Mínimo	Máximo	Médio
Nº de participações	6	16	9,58
Nº de vitórias	1	10	3,43

Variável	Mínimo	Máximo	Médio
Taxa de vitórias	8,33%	87,50%	35,64%
Total proposto	-	-	R$ 397.178,65 / licitação
Total vencido	-	-	R$ 394.565,48 / licitação
Qtd de anos	1	7	3,26
Nº de licitações por ano	1,17	10	3,32
Nº de cidades	1	12	5,52

O grupo D, o último dos conjuntos analisados, contempla 141 empresas, sendo 17 já investigadas pelo TCE/PB, o que representa uma taxa de investigação de 12,06%.

Observando a Tabela 5.2.2.4, verifica-se que as empresas do grupo D possuem um número de participações intermediário aos dos outros conjuntos. Apesar de sua taxa de vitórias ser menor apenas do que a do grupo A, o valor do conjunto D pode ser considerado baixo.

As empresas do grupo D concorreram em licitações durante uma quantidade média de anos, possuindo uma quantidade mediana de certames participados por ano. Essas companhias concorreram em um número considerável de cidades diferentes.

Assim, os quatro grupos criados a partir da aplicação do k-means em conjunto com o PSO no *dataset* reduzido podem ser descritos de forma resumida da seguinte maneira:

- Grupo A: engloba as empresas que participam pouco de licitações, mas vencem muitos certames;
- Grupo B: composto pelas empresas com taxas de participação muito baixas e com taxas de vitórias praticamente nula;
- Grupo C: compreende as empresas com altas taxas de participação e com baixas taxas de vitórias;
- Grupo D: composto por empresas com taxas de participação medianas e com baixas taxas de vitórias.

Assim como verificado no resultado obtido utilizando 3 clusters, os grupos A e C gerados pela clusterização dos dados adotando 4 conjuntos apresentam comportamentos bastante suspeitos. O grupo B, de modo semelhante ao verificado anteriormente, aparenta ser menos duvidoso, mas não pode ser considerado totalmente idôneo.

O grupo de empresas D, que foi obtido com a inclusão de mais um conjunto no processo de clusterização, pode ser considerado, dentre todos os grupos mencionados e

devido ao seu padrão de comportamento, como o conjunto menos suspeito. Apesar de sua taxa média de vitórias ser menor do que poderia ser esperado, deve-se considerar que o cometimento de fraudes nas licitações faz com que a concorrência real não siga exatamente o que seria (teoricamente) esperado.

Salienta-se que uma empresa não pode ser considerada culpada apenas por fazer parte de um dos conjuntos destacados. É preciso que uma investigação mais aprofundada seja realizada para determinar a culpabilidade ou não da empresa. Apesar disso, a classificação das companhias nos grupos gerados a partir do seu padrão de comportamento fornece uma boa noção geral da empresa e pode contribuir para os processos investigativos.

5.3. RESULTADOS FINAIS

Nas seções anteriores foram apresentados os resultados das duas metodologias empregadas neste trabalho para auxiliar o processo investigativo de fraudes em licitações. Nesta seção, será mostrado como tais resultados podem ser relacionados para melhorar as análises a serem realizadas.

O primeiro resultado apresentado utilizou conceitos da teoria dos grafos para gerar representações das relações que empresas mantiveram ao longo do período de tempo analisado e considerando os filtros que foram aplicados nos dados.

Já o segundo resultado obtido neste trabalho foi fruto da clusterização dos dados, o que possibilitou que as empresas analisadas pudessem ser divididas em diferentes grupos com base em suas características de participação nos certames estudados.

Unindo ambos os resultados alcançados, foi possível utilizar a teoria dos grafos para filtrar relações entre empresas e, a partir dos grupos estabelecidos na clusterização, caracterizar o comportamento de cada uma das companhias analisadas.

Destaca-se que na seção 5.1 foram apresentadas representações gráficas geradas a partir do *dataset* fornecido pelo TCE/PB e pré-processado com base nos critérios já mencionados. Contudo, como na seção 5.2 foi utilizado apenas uma parte desse conjunto de dados para a divisão das empresas em grupos, os grafos utilizados para exemplificar a união das duas técnicas foram gerados utilizando o *dataset* reduzido.

Logo, considerando o cenário de aplicação da clusterização usando três grupos e utilizando o *dataset* reduzido para a formação dos grafos, foi possível elaborar a

representação gráfica para todas as empresas analisadas, sem a aplicação de nenhum filtro. O grafo gerado pode ser visualizado na Figura 5.3.1.

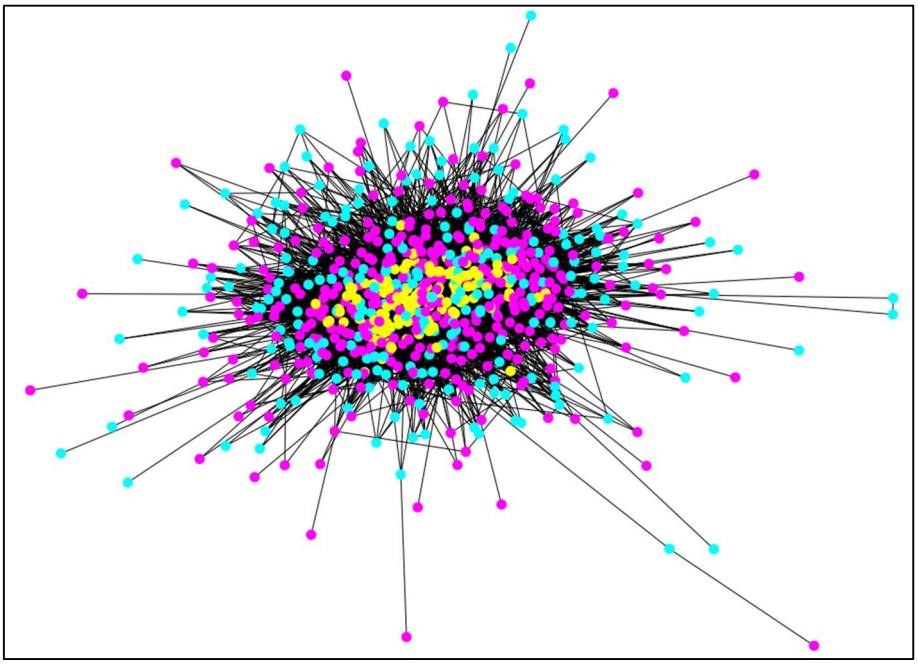


Figura 5.3.1 – Grafo gerado utilizando o *dataset* reduzido considerando 3 clusters.

O grafo gerado usando o *dataset* reduzido representa 728 empresas (nós), que mantiveram 13.773 relações entre si (arestas) durante os 8 anos de estudo e considerando os filtros aplicados para a diminuição do conjunto de dados.

No grafo mostrado na Figura 5.3.1, os nós estão coloridos em três cores, magenta, ciano e amarelo, que representam respectivamente os grupos A (empresas que participam de poucas licitações, mas ganham muito), B (companhias que participam de poucos certames e não ganham nenhum) e C (empresas que participam muito, mas ganham poucas licitações).

Para melhorar a análise do grafo mostrado na Figura 5.3.1, sua clique máxima foi obtida. Essa estrutura encontra-se exposta na Figura 5.3.2.

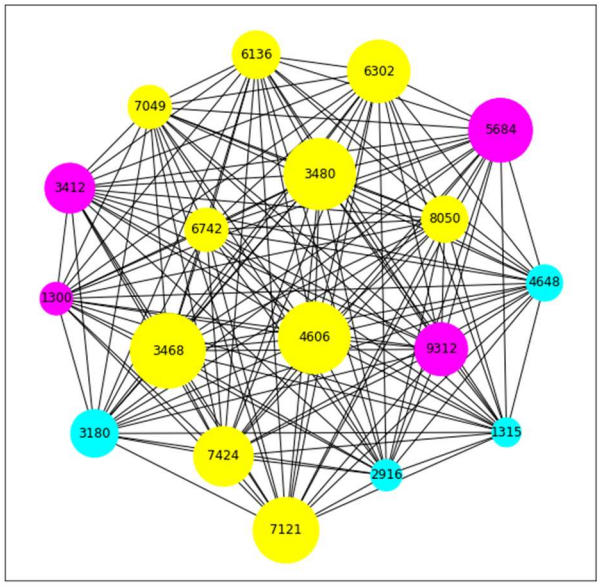


Figura 5.3.2 – Clique máxima do grafo gerado utilizando o *dataset* reduzido considerando 3 clusters.

A clique máxima obtida e mostrada na Figura 5.3.2 é formada por 18 nós, que representam as empresas 1300, 1315, 2916, 3180, 3412, 3468, 3480, 4606, 4648, 5684, 6136, 6302, 6742, 7049, 7121, 7424, 8050 e 9312. Nenhuma dessas empresas já foi alvo de processo investigativo movido pelo TCE/PB. Dentre os dezoito nós da clique máxima obtida, quatro pertencem ao grupo A, quatro ao conjunto B e dez à classe C.

Ou seja, do grupo destacado pelo algoritmo devido ao grau de relação entre os nós, quatro empresas participaram pouco de licitações, mas tiveram uma alta taxa de vitórias, quatro participaram pouco e não ganharam nenhum certame e dez participaram muito, mas ganharam muito poucas licitações.

Expandindo a clique máxima para a obtenção da clique quase máxima, foram adicionados 10 nós à estrutura. O resultado obtido é exibido na Figura 5.3.3.

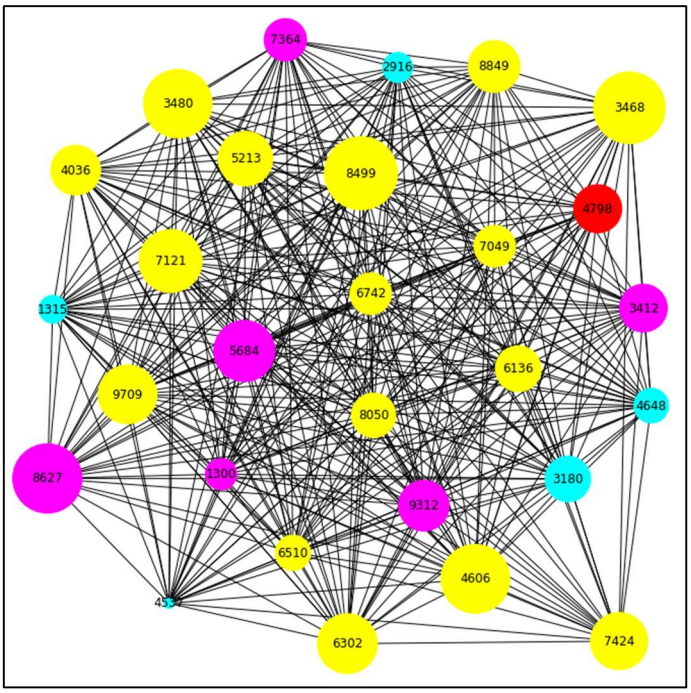


Figura 5.3.3 – Clique quase máxima do grafo gerado utilizando o *dataset* reduzido considerando 3 clusters.

Foram adicionados à clique máxima os nós das empresas 4036, 5213, 4798, 4537, 6510, 7364, 8499, 8849, 8627 e 9709, de modo a obter-se uma densidade de 0,97 para a clique quase máxima. Analisando a Figura 5.3.3, nota-se a presença de um nó vermelho, relativo à empresa 4798.

A coloração vermelha atribuída a tal nó indica que a empresa representada já foi alvo de processo investigativo pelo TCE/PB. Originalmente, a companhia pertencia ao grupo A, o de cor magenta.

Analisando a clique quase máxima mostrada na Figura 5.3.3, nota-se uma prevalência de nós amarelos, relacionados ao grupo de empresas que participam de muitas licitações, mas tem uma baixa taxa de vitórias. Comparando a quantidade dessas companhias com a das empresas que participam pouco de certames, mas ganham muito, tem-se uma relação de uma empresa do grupo A para cada 2,28 do C (considerando o nó 4798).

Isto é, para cada empresa da clique quase máxima gerada que apresenta as características de participar de poucas licitações e de ter uma alta taxa de vitórias, há 2 empresas que participam muito de certames, mas que ganham muito poucos.

Essa é, sem dúvidas, uma análise que pode ser realizada ao investigar-se a formação de conluios ou a ocorrência de outros tipos de fraudes em licitações. Salienta-se, contudo,

que outros indícios podem e devem ser utilizados para a obtenção de resultados melhores e mais assertivos.

O procedimento apresentado foi repetido, porém utilizando o resultado da clusterização que dividiu os dados em 4 grupos. Destaca-se que os grafos gerados são os mesmos para os dois cenários, mudando apenas a classificação que as empresas apresentam.

Com relação à coloração dos nós, manteve-se um padrão semelhante ao do resultado para 3 conjuntos de empresas, adicionando-se apenas a cor laranja para o quarto grupo. Logo, o grupo A (participa pouco, mas ganha muito) é representado pela cor magenta, o B (participa pouco e não ganha nada) pela cor ciano, o C (participa muito, mas ganha pouco) pelo amarelo e o D (tem participação média, mas ganha pouco) pelo laranja.

Dessa forma, o grafo contendo todas as relações do *dataset* reduzido gerado utilizando o padrão de coloração considerando 4 conjuntos de empresas pode ser visto na Figura 5.3.4.

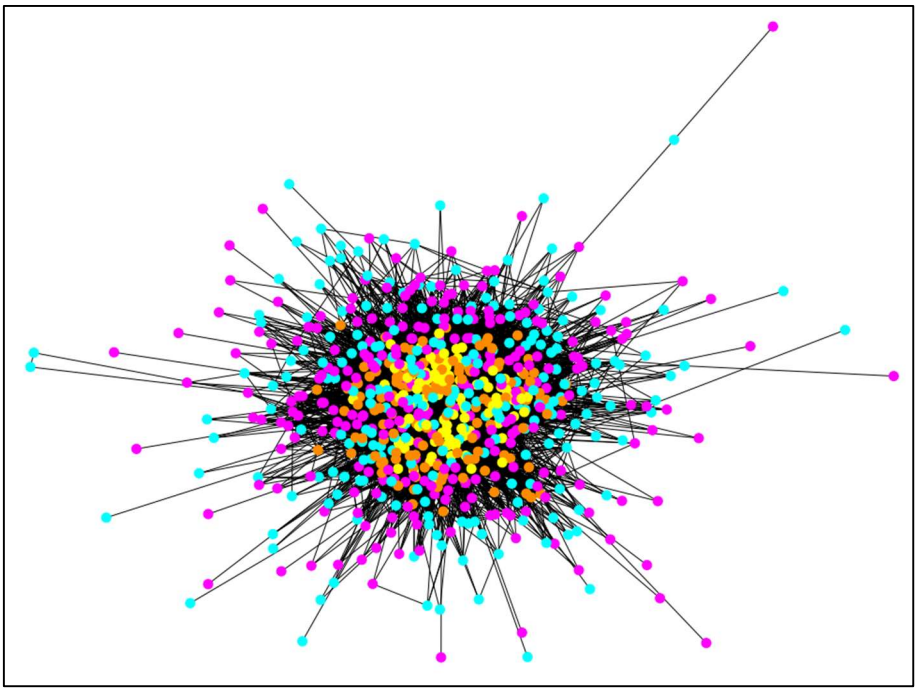


Figura 5.3.4 – Grafo gerado utilizando o *dataset* reduzido considerando 4 clusters.

Como dito anteriormente, o grafo exposto na Figura 5.3.4 apresenta 728 nós conectados por 13.773 arestas. Dessa estrutura foi retirada a clique máxima mostrada na Figura 5.3.5.

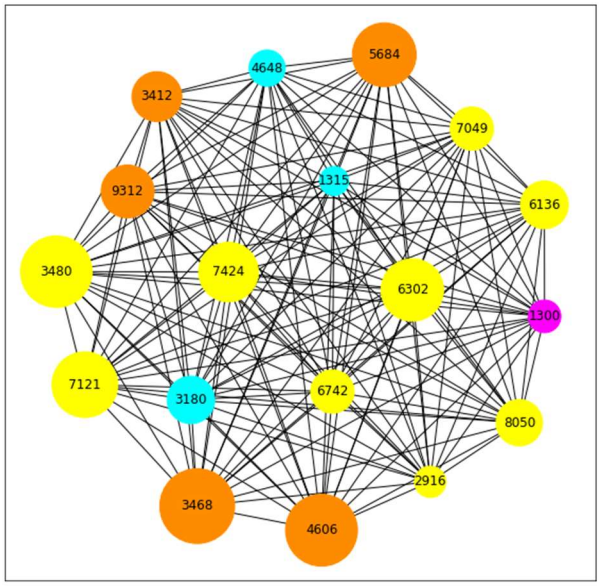


Figura 5.3.5 – Clique máxima do grafo gerado utilizando o *dataset* reduzido considerando 4 clusters.

Fazem parte da clique máxima mostrada na Figura 5.3.5 as empresas 1300, 1315, 2916, 3180, 3412, 3468, 3480, 4606, 4648, 5684, 6136, 6302, 6742, 7049, 7121, 7424, 8050, 9312. Os tamanhos dos nós relacionam-se com a taxa de vitórias das empresas representadas e as suas cores são referentes à classificação obtida pela aplicação da clusterização.

O panorama de classificação das empresas ao utilizar-se o cenário de 4 grupos é diferente do anterior, que considera apenas 3 conjuntos de empresas. Nesse resultado, tem-se que uma empresa participa de poucas licitações, mas tem alta taxa de vitórias, três participam pouco e não ganham nada, nove companhias participam muito, mas ganham muito pouco e cinco tem taxa de participação média e taxa de vitórias baixa.

Assim como no caso anterior, uma clique quase máxima do grafo da Figura 5.3.4 foi obtida a partir da adição de 10 nós à clique máxima mostrada na Figura 5.3.5. O subgrafo gerado a partir desse procedimento pode ser visualizado na Figura 5.3.6.

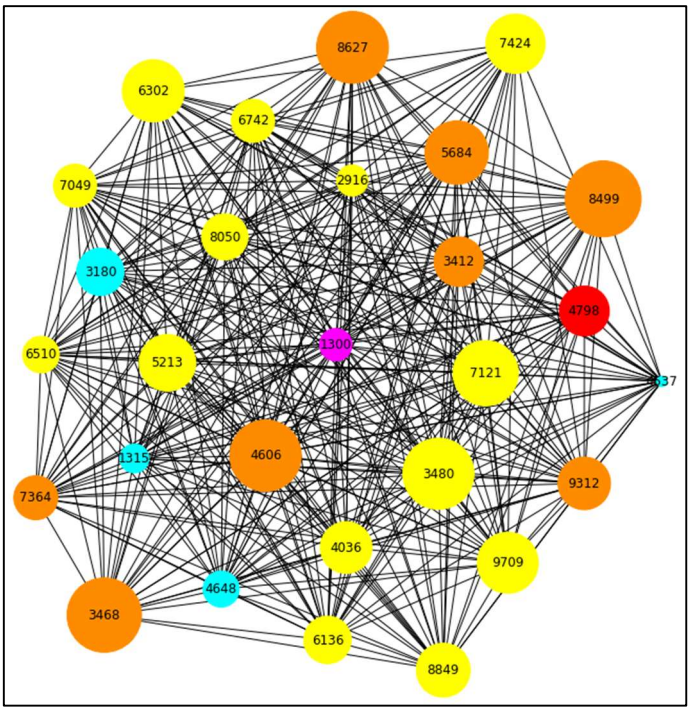


Figura 5.3.6 – Clique quase máxima do grafo gerado utilizando o *dataset* reduzido considerando 4 clusters.

A clique quase máxima exposta na Figura 5.3.6 possui densidade igual à 0,97. Verifica-se novamente a presença de uma empresa investigada no subgrafo, cujo código é 4798. Essa companhia foi classificada no grupo D, de coloração laranja, que representa as empresas que possuem participação média em licitações, mas apresentam taxa de vitórias relativamente baixa.

Comparando-se novamente a relação entre a quantidade de empresas que fazem parte do grupo das companhias que participam pouco, mas ganham muito, com as que participam muito, mas ganham pouco, constata-se que a relação que antes era de 1 para 2,28, ao utilizar-se 4 clusters passa para 1 para 15.

Ou seja, no subgrafo destacado, para que uma empresa participe pouco e apresente altas taxas de vitórias, é necessário que outras 15 participem muito e ganhem pouco. Salienta-se que esse valor de relação não é, necessariamente, grande. Uma maior quantidade de cenários tem que ser avaliados para se determinar os parâmetros do que é esperado e o que é estranho.

Como pode-se ver nos exemplos mostrados, a utilização dos grupos identificados utilizando a clusterização em conjunto com as relações evidenciadas pelos grafos permitiu uma análise mais minuciosa sobre as licitações, de modo a facilitar o reconhecimento de possíveis fraudes e possibilitar um melhor entendimento dos cenários investigados.

CAPÍTULO VI

6. DISCUSSÕES

Como visto, a metodologia desenvolvida neste trabalho utilizou a teoria de grafos e a clusterização de dados para identificar relações suspeitas que empresas mantiveram em processos licitatórios.

Para a exemplificação das técnicas propostas, foram utilizados dados de licitações públicas que ocorreram no estado da Paraíba entre 2014 e 2021, fornecidos pelo Tribunal de Contas do Estado da Paraíba (TCE/PB) por meio do portal Sagres Online.

Utilizando esses dados, foram exemplificados uma série de estudos que podem ser realizados. A partir dessas aplicações, verificou-se que nas cliques máximas e quase máximas destacadas nesta dissertação, 70 diferentes empresas foram evidenciadas devido ao seu padrão de comportamento. Dentre essas, as companhias 8050, 7049, 6136 e 4036 chamam atenção devido ao seu comportamento.

As empresas destacadas foram presenças constantes nas cliques obtidas do grafo que representa todos os dados estudados, mostrado na Figura 5.1.1, sendo que elas estiveram presentes em todas as cliques quase máximas obtidas, em todos os filtros analisados, com exceção do que selecionou apenas as empresas que já foram investigadas.

Isso mostra que tais companhias mantém relações muito próximas entre si e que apresentam um comportamento, de certa forma, suspeito, mesmo nenhuma delas tendo sido investigada pelo TCE/PB até então.

Considerando que todas essas empresas foram classificadas após o processo de clusterização dos dados no grupo C, tanto no cenário que considera 3 clusters quanto o que utiliza 4 grupos, o comportamento do grupo fica ainda mais suspeito. O grupo C representa empresas que participaram de muitas licitações, mas que ganharam muito poucos certames.

A suspeição de tais empresas deverá ser confirmada a partir de outros indícios de ocorrência de crime. O modelo proposto também pode ser melhorado, a partir da utilização

de um dataset mais completo e do uso de mais variáveis para a clusterização. Mesmo assim, os resultados preliminares são bastante promissores.

Os grafos exibidos nesta dissertação apresentaram alguns comportamentos interessantes. O primeiro deles está relacionado à aplicação do filtro nos dados que limitou o número de participações das empresas. Verificou-se que quanto mais restritivo esse filtro era, menor era a relação entre o número de nós dos grafos e entre a sua quantidade de arestas.

Isto é, ao selecionar-se apenas as companhias com alta participação em certames, o número de empresas restantes decaia em uma proporção bem maior do que a quantidade de relações que essas companhias mantinham entre si. Dessa forma, constatou-se que o grupo das empresas que possuem maior participação em licitações é um conjunto denso, com alto grau de relação.

O resultado obtido a partir da análise apenas das empresas que já foram investigadas pela TCE/PB também foi bastante interessante. Após a identificação da clique quase máxima desse cenário, foi verificado um conjunto de empresas que podem ser consideradas suspeitas de cometerem irregularidades em licitações, devido ao alto grau de relações entre as companhias e a grande presença de empresas que já foram alvo de processos investigativos.

Uma última verificação realizada a partir das visualizações gráficas está relacionada ao tipo de atuação das empresas. Apesar de a aquisição de materiais e o aluguel de equipamentos estar relacionado à construção civil, as empresas que geralmente realizam esses serviços não são as mesmas que realizam outras atividades de engenharia civil, como a construção e/ou a reforma de obras e pavimentação de vias.

Sendo assim, trabalhos futuros que venham a estudar licitações da construção civil afim de verificar relações suspeitas entre empresas podem separar da análise os certames voltados para a aquisição de materiais e de equipamentos dos demais. Dessa forma, espera-se que os resultados obtidos sejam mais precisos, devido à diminuição da variância dos dados.

Analisando os resultados das divisões das empresas realizada pela clusterização dos dados, verifica-se a definição de grupos com características bem definidas. O conjunto que contempla as empresas com baixa taxa de participação e uma taxa de vitórias praticamente nula é bastante curioso.

Avaliando as características desse conjunto, é possível supor que ele contempla empresas que após participarem de alguns poucos processos licitatórios e não obterem êxito, acabam desistindo de concorrer em certames. Contudo, essa suposição não é,

necessariamente, unânime para todo o grupo e nem isenta as empresas da prática de irregularidades.

Outro grupo destacado pelo agrupamento das empresas é composto por companhias com baixo número de participações, mas elevada taxa de vitórias. Isto é, são empresas que participam de poucas licitações, mas quando competem, possuem grande chance de vencer. Esse comportamento é bastante incomum e levanta alguns questionamentos quanto a sua legalidade.

Algumas das empresas estudadas foram agrupadas em um conjunto onde os números de participação em licitações são bastante elevados, mas as taxas de vitórias são baixas. Trata-se de um comportamento contrário ao do grupo destacado anteriormente. Apesar disso, o grupo das empresas que perdem grande parte das muitas licitações em que participam é bastante suspeito.

Nota-se, ainda, que a adição de um cluster à divisão dos dados criou o grupo das empresas intermediárias, que possuem um número de participações médio e uma taxa de vitórias relativamente abaixo do que poderia ser esperado. Mesmo assim, dentre todos os conjuntos identificados neste trabalho, tal grupo pode ser considerado como um dos menos suspeitos.

Vale salientar que até então as investigações movidas por órgãos públicos, a exemplo do TCE/PB, eram realizadas sem nenhum padrão claramente definido. Assim sendo, as taxas de investigação verificadas para cada um dos grupos de empresas identificados podem não representar verdadeiramente a suspeição dos conjuntos.

Destaca-se, por fim, que a metodologia proposta neste trabalho se mostrou bastante viável, apresentando excelentes resultados. Trata-se de um método fácil de ser aplicado e que apresenta produtos com alto potencial de utilização.

CAPÍTULO VII

7. CONCLUSÃO

O objetivo deste trabalho foi sugerir uma nova metodologia para auxiliar o processo investigativo de licitações. Para isso, foram propostas as utilizações de técnicas da teoria dos grafos em conjunto com algoritmos de clusterização de dados para a identificação de relações suspeitas entre empresas.

Utilizando a teoria dos grafos, as relações que as empresas mantiveram entre si nas licitações trabalhadas puderam ser visualizadas, identificadas e analisadas com maior detalhamento e precisão.

Já com o auxílio do algoritmo k-means combinado com o uso do PSO, as empresas que participaram dos processos licitatórios estudados foram divididas em grupos com base em seu padrão de comportamento nos certames analisados.

Unindo os resultados das duas técnicas empregadas, foi possível identificar grupos de empresas suspeitos de cometerem irregularidades em processos licitatórios, devido às relações mantidas entre as companhias e às suas formas de atuação.

Neste trabalho, ficou evidente a importância da utilização de informações que indiquem se uma determinada empresa já foi investigada ou se ela pode ter cometido fraudes. Com a utilização desses elementos, a identificação de novos suspeitos pode ser realizada de forma mais assertiva.

Além disso, ficou claramente demonstrada a necessidade de que dados de licitações sejam divulgadas pelos órgãos públicos e que o cadastramento das informações seja realizado de forma correta e padronizada.

Nesta dissertação foram utilizados dados de licitações públicas ocorridas no Estado da Paraíba entre 2014 e 2021 e que foram fornecidos pelo Portal Sagres Online. Apesar das incoerências dos dados fornecidos pelo TCE/PB que dificultaram a aplicação do modelo proposto, destaca-se a grande importância que a base de dados do Sagres possui.

Como visto na revisão bibliográfica realizada neste trabalho, apesar de muitos métodos terem sido desenvolvidos visando a identificação de possíveis irregularidades em licitações, não há nenhuma técnica consolidada que seja utilizada para tal fim.

Destaca-se que a metodologia proposta neste trabalho apresentou excelentes resultados. Trata-se de um método simples, de fácil implementação e execução cujos produtos podem contribuir para investigações criminais. Além disso, devido à sua rapidez, as técnicas apresentadas podem ser constantemente aplicadas e adaptadas para novos dados.

A metodologia sugerida também apresenta o diferencial de não exigir um *dataset* com parâmetros fixos. As variáveis utilizadas nas análises podem ser modificadas e adaptadas considerando as informações disponíveis.

Por fim, outra vantagem do método sugerido é que ele pode ser empregado tanto antes da realização da contratação da empresa, ainda durante o processo licitatório, evitando-se, assim, a consolidação do crime, quanto durante um processo investigativo, afim de obter-se indícios da ocorrência de fraudes.

Salienta-se que as técnicas propostas neste trabalho são apenas sugestões de métodos que podem ser utilizados para auxiliar as investigações de formação de conluio. Assim, a metodologia sugerida pode ser empregada de forma isolada ou em conjunto com outras técnicas, visando a obtenção de resultados mais precisos.

Ademais, os grafos utilizados neste trabalho foram gerados usando as relações que as empresas mantiveram durante processos licitatórios. Entretanto, outros tipos de informações também poderiam ser usados como, por exemplo, dados de movimentações bancárias ou de ligações telefônicas.

Dentre os possíveis trabalhos futuros, sugere-se que metodologia desenvolvida neste trabalho seja aplicada em outros *datasets*, afim de avaliar como cada conjunto de dados se comporta.

Com relação à aplicação da clusterização, novas análises podem ser realizadas utilizando um *dataset* com mais informações. Para isso, contudo, será necessário garantir que todos os dados relacionados aos custos das licitações estão preenchidos corretamente. Por fim, pode-se realizar a clusterização utilizando-se outros tipos ou um maior número de variáveis, visando uma identificação mais assertiva dos padrões de comportamento das empresas.

AGRADECIMENTOS

Agradeço a concessão da bolsa de mestrado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) no âmbito do Programa de Cooperação Acadêmica (PROCAD) com parceria entre a Polícia Federal e a Universidade Federal da Paraíba (UFPB).

REFERÊNCIAS

BECCENERI, J. C., 2008. *Meta-heurísticas e Otimização Combinatória: Aplicações em Problemas Ambientais*. INPE, Sao José dos Campos.

BRASIL, 2018. *Decreto nº 9.412, de 18 de junho de 2018, atualiza os valores das modalidades de licitação de que trata o art. 23 da lei nº 8.666, de 21 de junho de 1993.*

BRASIL, 2021. *Lei nº 14.133, de 1 de abril de 2021.* Lei de Licitações e Contratos Administrativos. Diário Oficial da República Federativa do Brasil, Brasília, DF, 1 abril 2021.

CARVALHO, M. A. M., 2020. *BCC204 – Teoria dos Grafos: Aula 01.* Ouro Preto. Disponível em: <http://www.decom.ufop.br/marco/site_media/uploads/bcc204/01_aula_01.pdf>. Acesso em fevereiro de 2023.

COELHO FILHO, O. P., MARTINHON, C. A., CABRAL, L. D. A. F., 2013. *Uma Abordagem Melhorada do Algoritmo de Otimização por Enxame de Partículas para o Problema de Clusterização de Dados.* Simpósio Brasileiro de Pesquisa Operacional. Natal:[sn], p. 2135-2146.

CUIABANO, S. M., LEANDRO, T., OLIVEIRA, G. A. S., BOGOSSIAN, P., 2014. *Filtrando cartéis: a contribuição da literatura econômica na identificação de comportamentos colusivos*. Revista de Defesa da Concorrência, 2(2), 43-63.

DA UNIÃO, Brasil Tribunal de Contas. *Licitações e contratos: orientações e jurisprudência do TCU*. 2010.

DORIGO, M., DE OCA, M. A. M., ENGELBRECHT, A., 2008. *Particle swarm optimization.* Scholarpedia, v. 3, n. 11, p. 1486.

DOS SANTOS, T. D., JUSTEL, C. M., 2015. *Alguns experimentos com a conectividade algébrica em grafos aleatórios.* Ciência e Tecnologia, p. 48.

FOREMNY, A., ANYSZ, H., 2018. *The collusion detection in public procurements–selected methods applied for the road construction industry in Poland.* In: MATEC Web of Conferences. EDP Sciences. p. 04002.

FRAGA, A. A., 2017. *Detecção de casos suspeitos de fraudes em licitações realizadas nos municípios da Paraíba: uma aplicação de técnicas de mineração de dados.* Dissertação de M. Sc., UFPB, João Pessoa, PB, Brasil.

FRIEDMAN, L., 1956. *A competitive-bidding strategy.* Oper. Res. 4 (1): 104–112.

HAGBERG, A., SWART, P., CHULT, D. S., 2008. *Exploring network structure, dynamics, and function using NetworkX*. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

KENNEDY, J., EBERHART, R., 1995. *Particle swarm optimization.* In: Proceedings of ICNN'95-international conference on neural networks. IEEE. p. 1942-1948.

LIMA, M. C., 2010. *Comparação de custos referenciais do DNIT e licitações bem sucedidas.* Revista do TCU, n. 118, p. 61-66.

LIMA, M. C., 2021. *Deep Vacuity: detecção e classificação automática de padrões com risco de conluio em dados públicos de licitações de obras.* Dissertação de M. Sc., UNB, Brasília, DF, Brasil.

MODRUŠAN, N., RABUZIN, K., MRŠIC, L., 2021. *Review of Public Procurement Fraud Detection Techniques Powered by Emerging Technologies*. International Journal of Advanced Computer Science and Applications, v. 12, n. 2.

NEVES, C. O. M., 2016. *Caracterização das áreas queimadas no estado do Tocantins no ano de 2014.* Dissertação de M. Sc., UFT, Gurupi, TO, Brasil.

NOGUEIRA JÚNIOR, D. C., 2017. *Grafos e problemas de caminhos*. Dissertação de M. Sc., UFV, Viçosa, MG, Brasil.

O'SULLIVAN, A., SHEFFRIN, S. M., 2003. *Economics: Principles in action*.

PATEL, G. K., DABHI, V. K. and PRAJAPATI, H. B., 2017. *Clustering Using a Combination of Particle Swarm Optimization and K-means.* Journal of Intelligent Systems, vol. 26, no. 3, pp. 457-469. https://doi.org/10.1515/jisys-2015-0099

PEDNEKAR, A. M., 2019. *Optimal initialization of K-means using Particle Swarm Optimization.* arXiv preprint arXiv:1904.09098.

PNUD, 2022. *Human development report 2021/2022.* United Nations Development Programme. Technical report.

PORTAL DA TRANSPARÊNCIA DO GOVERNO DO BRASIL, 2022. *Licitações com contratação realizada.* Disponível em: <https://portaldatransparencia.gov.br/licitacoes?ano=2022>. Acesso em janeiro de 2023.

PRESTES, E., 2016. *Introdução à Teoria dos Grafos*. Universidade Federal do Rio Grande do Sul, Instituto de Informática, Departamento de Informática Teórica, Tech. Rep.

RINGNÉR, M., 2008. *What is principal component analysis?* Nature Biotechnology, 26(3), 303–304. https://doi.org/10.1038/nbt0308-303

RODRÍGUEZ, M. J. G., RODRÍGUEZ-MONTEQUÍN, V., BALLESTEROS-PÉRES, P., LOVE, P. E. D., SIGNOR, R., 2022. *Collusion detection in public procurement*

auctions with machine learning algorithms. Automation in Construction, v. 133, p. 104047.

SIGNOR, R., LOVE, P. E. D., OLIVERIA JR, A., LOPES, A. O., OLIVEIRA JR, P. S. *Public infrastructure procurement: detecting collusion in capped first-priced auctions*. Journal of Infrastructure Systems, v. 26, n. 2, p. 05020002, 2020.

SMUDA, F., 2013. *Cartel overcharges and the deterrent effect of eu competition law.* Journal of Competition Law and Economics, 10(1), 63–86. https://doi.org/10.1093/joclec/nht012

TCE/PB, 2010. *Sagres online: um instrumento de controle social.* Tribunal de Contas do Estado da Paraíba. 58 p. João Pessoa: A União.

TCE/SC. *Sistema de Fiscalização Integrada de Gestão - e-Sfinge*. Disponível em: <https://www.tcesc.tc.br/esfinge/informacoes>. Acesso em agosto de 2022.

VELASCO, R. B., CARPANESE, I., INTERIAN, R., PAULO NETO, O. C., RIBEIRO, C. C., 2021. *A decision support system for fraud detection in public procurement*. International Transactions in Operational Research, *28*(1), 27-47.

ZHU, A., 2022. *7 Most Asked Questions on K-Means Clustering*. Towards Data Science. Disponível em: <https://towardsdatascience.com/explain-ml-in-a-simple-way-k-means-clustering-e925d019743b>. Acesso em fevereiro de 2023.

ZÜGE, A. P., 2017. *Algoritmos para o problema da clique máxima: análise e comparação experimental.* Tese de D.Sc., UFPR, Curitiba, PR, Brasil.