CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

ANÁLISE DE GRAFOS PARA APOIO EM AUDITORIA DE LICITAÇÕES PÚBLICAS

Wellington Souza Amaral

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Orientador(a): Leonardo Silva de Lima
Coorientador(a): Eduardo Bezerra

Rio de Janeiro,

Janeiro de 2020

ANÁLISE DE GRAFOS PARA APOIO EM AUDITORIA DE LICITAÇÕES PÚBLICAS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Wellington Souza Amaral

Banca Examinadora:

Presidente, Professor D.Sc. Leonardo Silva de Lima(UFPR) (Orientador(a))

Professor D.Sc. Eduardo Bezerra(CEFET-RJ) (Coorientador(a))

Professor D.Sc. Eduardo Soares Ogasawara (CEFET-RJ)

Professora D.Sc. Claudia Marcela Justel (IME)

Rio de Janeiro,

Janeiro de 2020

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

A485 Amaral, Wellington Souza.
Análise de grafos para apoio em auditoria de licitações públicas / Wellington Souza Amaral – 2020.
95f. : il. (algumas color.)., grafs., tabs. ; enc.

Dissertação (Mestrado). Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2020.
Bibliografia: f. 92-95.
Orientador: Leonardo Silva de Lima.
Coorientador: Eduardo Bezerra da Silva.

1. Teoria dos grafos. 2. Licitação pública. 3. Auditoria. I. Lima, Leonardo Silva de (Orient.). II. Silva, Eduardo Bezerra da (Coorient.). III. Título.

CDD 511.5

Elaborada pela bibliotecária Vanessa Suane de Souza CRB-7/6753

DEDICATÓRIA

A Júlia, Nathan, Nicole e Eloah, que eu possa ajudar a fazer o Brasil de vocês ainda melhor do que meu.

AGRADECIMENTOS

Em primeiro lugar aos meus pais, Elton e Ivete, pelo amor e carinho incondicional que moldaram o homem que eu sou.

A toda a minha família por me apoiar e acreditar em mim antes mesmo que eu acreditasse.

Aos amigos da Secretaria Geral do Controle Externo pelo incentivo, especial a Talita Dourado, Alberto Tavares, Bruno Mattos e Marcos Ferreira.

Aos professores do programa pela dedicação ao ensino e à pesquisa.

Ao companheiro Sérgio Lino pela sua ótima contribuição com a revisão textual.

Aos novos amigos do PPCIC e PPPRO que tornaram essa jornada mais fácil, especialmente o grande amigo Alexandre Cunha.

RESUMO

Análise de grafos para apoio em auditoria de licitações públicas

O presente trabalho desenvolve uma metodologia para identificar não conformidades nos processos licitatórios realizados por órgãos do estado do Rio de Janeiro. Nosso interesse é motivado pela necessidade de selecionar gastos públicos suspeitos de conter irregularidades, já que é impossível investigar detalhadamente todos os gastos estaduais e contratos públicos. São utilizados métodos relacionados para mineração de dados, teoria de grafos e teoria da informação. O método adotado é modelar o problema em três tipos de redes: uma rede bipartida de empresas e órgãos públicos; uma segunda rede composta apenas por empresas, e uma terceira rede formada por empresas e órgãos públicos, consolidando os resultados obtidos nas duas redes anteriores. Métodos gráficos de mineração foram aplicados a todas essas redes para identificar conluio. Os experimentos computacionais foram realizados em um conjunto de dados real com mais de 120 redes. Nossos resultados apontaram para alguns órgãos públicos que deveriam ter seus contratos investigados.

Palavras-chave: Detecção de comunidades; mineração em grafos; detecção de irregularidades; entropia; mineração de dados; compras públicas; detecção de conluio

ABSTRACT

Analysis of graphs for support in auditing public bids

The present work develops a methodology for identifying non-conformities in the bidding processes carried out by agencies of the Rio de Janeiro state. Our interest is motivated by the need to select public expenditures most suspected of containing irregularities, as it is impossible to investigate in detail all state expenditures and public contracts. Related methods for data mining, graph theory, and information theory are used. The method adopted is to model the problem into three types of networks: a bipartite network of companies and public agencies; a second network composed only of companies, and a third network formed by companies and public agencies consolidating the results obtained in the previous two networks. Mining graph methods were applied to all of these networks in order to identify collusion. The computational experiments were performed over a real dataset with more than 120 networks. Our results pointed out to some public agencies that should have its contracts investigated.

Keywords:Community detection; graph mining; irregularity detection; entropy; data mining; Government procurement auctions; collusion detection.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS E SIGLAS

CNAE	Classificação Nacional De Atividade Econômica
CNM	Clauset, Newman E Moore
CRISP-DM	Cross Industry Standard Process For Data Mining
GN	Girván-Newman
IBGE	Instituto Brasileiro De Geografia Estatística
K-S	Kolmogorov-Smirnov
PIB	Produto Interno Bruto
TCE-RJ	Tribunal De Contas Do Estado Do Rio De Janeiro

SUMÁRIO

1- Introdução

1.1- Contextualização

Os órgãos de controle são responsáveis pela fiscalização de recursos públicos [Brasil, 1988]. Estes órgãos não têm recursos econômicos e humanos para uma análise pormenorizada de toda despesa ou ato realizado pelos órgãos controlados [Balaniuk, 2010]. Em sua estratégia de controle, tais órgãos se utilizam de alguns critérios para selecionar quais atos e despesas serão analisadas em profundidade [TCU, 2016]. A mineração de dados tem se mostrado uma importante ferramenta para a realização de análises preliminares sobre todo o universo passível de controle [Balaniuk, 2010].

Os órgãos de controle da administração pública, como os Tribunais de Contas e os órgãos de controle interno, são responsáveis por fiscalizar a atuação das unidades administrativas dos órgãos públicos, sendo inúmeros os tipos de atos e despesas passíveis de ações de controle. Entre os tipos de atos passíveis de fiscalizações, há um grande interesse no controle de contratos públicos. Esses contratos são celebrados entre a administração pública e particulares, para a prestação de serviços de acordo com o interesse público.

Os órgãos de controle não dispõem de recursos humanos e econômicos suficientes para a fiscalização de todos os contratos públicos. Tome o exemplo do Tribunal de Contas do Estado do Rio de Janeiro (TCE-RJ), órgão com previsão constitucional que é responsável por auditar qualquer despesa ou receita de todos os órgãos do Estado do Rio de Janeiro e de todos órgãos dos municípios do Rio de Janeiro, com a exceção da capital. O TCE-RJ possui 519 (quinhentos e dezenove) auditores, segundo dados do próprio órgão [TCE-RJ, 2014]. Esses auditores são responsáveis por auditar todas as contratações dos órgãos jurisdicionados ao TCE-RJ. Só o Estado do Rio de Janeiro, em 2018, gastou cerca de R$ 3,5 bilhões em contratação de bens e serviços para realizar as suas funções de Governo [TCE-RJ, 2016].

O número de auditores é insuficiente para a fiscalização pormenorizada de cada contratação. É econômica e tecnicamente inviável a fiscalização detalhada dos aspectos

de todas essas compras realizadas. Os auditores selecionam uma pequena parcela das compras para uma fiscalização pormenorizada, com vistas à identificação de irregularidades em meio a um universo bem maior de compras realizadas pelos órgãos públicos. Os auditores se valem de alguns critérios para seleção, como risco, relevância e materialidade [TCE-RJ, 2010].

Dado o enorme número de contratações passíveis de fiscalização, a atuação de maneira eficiente na identificação de irregularidades é um desafio aos órgãos de controle. A seleção de objetos e ações de controle é elemento-chave para a efetividade do controle na fiscalização desses recursos [TCU, 2016]. Questões relacionadas a desvios de recursos financeiros do setor público, nas suas mais variadas formas, inclusive resultantes de práticas anticompetitivas aliadas à corrupção de agentes governamentais, têm ganhado crescente interesse por parte da sociedade. O Brasil ocupa a 105º posição no ranking de 2018 que avalia a percepção da corrupção no setor público em 180 países. O índice de percepção da corrupção é desenvolvido pela organização não governamental Transparência Internacional. Quanto pior a posição no ranking mais o país é considerado corrupto de acordo com a percepção de sua população [Transparency-International, 2019].

Há vários exemplos recentes de fraudes em âmbito nacional e estadual investigadas pela Polícia Federal, a saber, Operação Vampiro (2004), Operação Sanguessuga (2006), Operação Carta Marcada (2006), a Operação Lava-Jato (2012) e seus desdobramentos em diversas fases e a Operação Pão Nosso (2018). Considerando-se o valor médio anual de R$ 300 bilhões para compras e aquisições diversas em processos licitatórios, estima-se que práticas fraudulentas no Brasil gerem um prejuízo em torno de R$ 25 a R$ 40 bilhões, o que representa uma média de 10% de desvio de recursos públicos, segundo dados da Secretaria de Defesa Econômica do Ministério da Justiça [Campos, 2008].

Apesar de sua importância na economia do setor público, os mercados de licitações são vulneráveis a ações fraudulentas e à corrupção. Isso decorre de suas características intrínsecas, que possibilitam diversas interações entre agentes públicos e privados. Por outro lado, Martins Junior and Braz [2010]; Balaniuk [2010] apontam a importância da mineração e análise de dados a partir das bases de dados disponíveis para produção de conhecimento dos Tribunais de Contas, em auxílio a sua missão. A estes órgãos cabe a fiscalização contábil, orçamentária, operacional e patrimonial dos

órgãos da administração direta e indireta [Brasil, 1988].

Os trabalhos de pesquisas de Melo and Ferreira [2016]; Balaniuk [2010] já provaram que o uso de métodos afetos à mineração de dados podem aumentar significativamente a capacidade analítica dos órgãos de controle e contribuir para a realização da sua atividade fim. O presente trabalho propõe uma metodologia a partir do uso de ferramentas da mineração de grafos com o propósito de auxiliar órgãos de controle fornecendo uma lista de órgãos e contratos, com possíveis irregularidades, para uma fiscalização em profundidade. Trabalhos de mineração de dados contribuem principalmente para a produção de conhecimento estratégico que subsidie a tomada de decisão, a escolha dos objetos de auditoria e dos instrumentos de fiscalização.

1.2- Motivação

Diante do exposto, pesquisas científicas que desenvolvam métodos para apoiar a auditoria de licitações públicas se mostram relevantes. Nos Tribunais de Contas, em particular, a estruturação e utilização de uma metodologia de análise e produção de conhecimento a partir de grandes bases de dados pode levar a ações muito mais eficazes no exercício da sua atividade finalística de controle externo [Balaniuk, 2010].

Trabalhos anteriores já investigaram o uso de métodos mineração de dados para aumentar a capacidade analítica dos órgãos de controle. As abordagens utilizadas para esse fim são diversas, como métodos estatísticos Carvalho et al. [2014]; Padhi and Mohapatra [2011], regras de associação Souza and Pereira [2009]; Ralha and Silva [2012]; Fraga et al. [2017], aprendizado supervisionado Arief et al. [2016], mineração de texto e análise de agrupamento Davydenko et al. [2017], análise de redes sociais Melo and Ferreira [2016], Probabilistic Ontology Web Language em Carvalho et al. [2013] e redes neurais artificiais Domingos et al. [2016]. O presente trabalho apresenta uma metodologia não supervisionada para auxiliar órgãos de controle na seleção de licitações para fiscalização realizada por esses órgãos. No capítulo 3 são detalhadas esses e outros trabalhos correlatos ao tema da presente pesquisa.

Neste estudo, desenvolvemos duas abordagens centradas na mineração de grafos. Na primeira abordagem, modelamos as licitações públicas, os órgão públicos e as

empresas em um grafo bipartido. Os nós representam os órgãos públicos e as empresas participantes de licitações. Uma empresa tem uma conexão com um órgão público se participou de alguma licitação realizada por aquele órgão. Note que, desta forma, não há conexões entre as empresas e nem entre os órgãos públicos. Calculamos a entropia de cada nó que representa um órgão público. A partir da medida de entropia, os órgãos são ordenados para possíveis avaliações pelo órgão de controle. Esta abordagem é inovadora quando aplicada a identificação de irregularidades em contratos públicos, uma vez que não há trabalhos em fraudes de licitações públicas que façam análises a partir da entropia. Esta metodologia foi aplicada sobre os dados de licitações públicas ocorridas entre 2010 e 2018 de diversos órgãos públicos do Estado do Rio de Janeiro.

Na segunda abordagem, construímos um grafo em que os nós representam todas as empresas que tenham participado de alguma licitação e as arestas conectam empresas que tenham participado da mesma licitação. Detectamos as comunidades formadas a fim de identificar grupos de empresas em conluio com o objetivo de simular uma concorrência. Em nossos experimentos, aplicamos a metodologia sobre os dados de licitações públicas entre 2010 e 2018 de diversos órgãos do Estado do Rio de Janeiro. Em seguida, um grafo juntando os resultados das duas abordagens anteriores é produzido e uma indicação de empresas que provavelmente possam agir em conluio é apresentada para o órgão fiscalizador.

1.3- Pergunta da pesquisa

O tema da pesquisa objetiva responder à seguinte questão:

“A partir de um conjunto de licitações realizadas entre órgãos governamentais e empresas privadas, quais os órgãos e quais os contratos deste órgão que devem ser selecionados para uma investigação mais detalhada em busca de possíveis irregularidades?”

1.4- Objetivo

REver objetivo O objetivo é fornecer uma metodologia que possa indicar uma lista de órgãos com possíveis indícios de irregularidades em contratos públicos a partir de uma modelagem por Teoria de Grafos.

O objetivo desta dissertação é fornecer uma metodologia que possa indicar uma lista de órgãos com possíveis indícios de irregularidades em contratos públicos a partir de técnicas da Teoria dos Grafos e aplicar mineração de dados em grafos. Essa metodologia é aplicada em particular ao Tribunal de Contas do Estado do Rio de Janeiro. Desta forma, vários objetivos específicos são almejados:

- Selecionar e processar os dados das compras públicas;
- Classificar as licitações conforme a atividade econômica em que estão inseridas;
- Medir e avaliar o valor de entropia, medida da Teoria da Informação, de cada órgão público;
- Identificar comunidades de empresas com indícios de ação em conluio para simular concorrência em licitações públicas.

1.5- Metodologia

A metodologia do trabalho considera as seguintes macro-etapas:

1. Levantamento bibliográfico dos principais artigos que abordaram o problema aqui proposto
2. Obtenção do banco de dados com todas as licitações públicas realizadas num dado período de tempo junto ao Tribunal de Contas do Estado do Rio de Janeiro ;
3. Manipulação/tratamento dos dados brutos;
4. Duas abordagens principais são desenvolvidas. Para a primeira abordagem, temos os seguintes métodos:

(a) Modelar o conjunto de dados como um grafo bipartido valorado, onde o peso nas arestas indica o número de licitações entre uma empresa e um órgão;

(b) Determinar o valor de entropia, medida da Teoria da Informação, de cada empresa e de cada órgão público representado;

(c) Ordenar os órgãos em ordem não-crescente e selecionar k órgãos com menor entropia como aqueles com maior risco de irregularidades.

Para a segunda abordagem temos os seguintes métodos:

(a) Modelar o conjunto de dados como um grafo valorado, onde os pesos nas arestas indicam os vínculos entre as empresas;

(b) Detectar as comunidades encontradas pelos métodos de Girvan-Newman (GN) e Clauset, Newman e Moore Clauset, Newman e Moore (CNM) no grafo de empresas. A partir deste resultado busca-se identificar comunidades com forte coesão obtidas por ambos algoritmos e indicar uma lista de empresas suspeitas de irregularidades;

(c) Construir um grafo de empresas e órgãos a partir dos resultados das duas abordagens acima, onde as comunidades de empresas com maior coesão encontradas na segunda abordagem são conectadas aos órgãos onde participaram de licitações. Uma lista de órgãos e contratos com possíveis irregularidades é sugerida.

A fim de efetivar a metodologia desenvolvida, foram implementadas as funções, rotinas e programas necessários em linguagem de programação Python. Esta metodologia foi aplicada sobre os dados de licitações públicas ocorridas entre 2010 e 2018 de diversos órgãos públicos do Estado do Rio de Janeiro.

É importante mencionar que o uso de abordagens tradicionais para mensurar os resultados de predição dos classificadores não é possível no presente trabalho, uma vez que: a amostra de dados não é totalmente rotulada; não é possível selecionar uma amostra do universo das contratações e realizar auditorias de fiscalização a fim de rotulá-las, uma vez que isso é economicamente inviável; não há rótulos adequados para a identificação de quais contratações são irregulares e que tipo de irregularidade ocorre; também não é possível iniciar um trabalho de rotulagem dos dados, já que seria necessário realizar várias auditorias e isso tem um custo proibitivo de realização. A

ausência dessas informações tem reflexo na metodologia desenvolvida neste trabalho que é discutida no Capítulo 4

1.6- Organização do Texto

Este trabalho está organizado nas seguintes seções: Capítulo 1, que introduz a temática, o problema de pesquisa e a metodologia utilizada na pesquisa; O Capítulo 2 apresenta temas, conceitos e técnicas utilizados para compreensão dos demais temas tratados nos capítulos seguintes. O Capítulo 3 apresenta outros trabalhos correlatos ao tema da pesquisa. No Capítulo 4 é apresentada a metodologia desenvolvida. O Capítulo 5 apresenta os resultados alcançados no conjunto de dados das contratações dos órgãos do Estado do Rio de Janeiro utilizando-se a metodologia apresentada no Capítulo 4. O Capítulo 6 apresenta as conclusões a respeito dos resultados e perspectivas de trabalhos futuros.

2- Referencial teórico

Neste capítulo, são apresentados alguns conceitos importantes que serão utilizados nos próximos capítulos e seções. A Seção 2.1 apresenta o contexto em que as técnicas de mineração de dados aqui apresentadas serão aplicadas; a Seção 2.2 apresenta a estrutura matemática em que serão modeladas as compras públicas; a Seção 2.3 discute uma medida importante que vai permitir avaliar e comparar a distribuição da sequência de participações das empresas em licitações de cada órgão; a Seção 2.4 apresenta um teste estatístico Kolmogorov-Smirnov (K-S) para verificar a aderência de uma distribuição estatística a um conjunto de dados; a Seção 2.5 apresenta uma ferramenta gráfica para representar a variação de dados observados e detectar outliers.

2.1- Domínio de negócio

Nesta seção são apresentados conceitos relativos ao domínio onde a metodologia de mineração de dados desenvolvida é aplicada. A Seção 2.1.1 explica como a contratação pública é realizada; a Seção 2.1.4 apresenta a atribuição do TCE-RJ, que é responsável, dentre outras atribuições, por auditar licitações públicas; a Seção 2.1.2 tipifica o que é irregularidade no contexto desta pesquisa; a Seção 2.1.3 apresenta como é estruturada a Classificação Nacional de Atividade Econômica (CNAE) e como ela será importante mais adiante na separação das empresas de acordo com o seu ramo de atividade econômica.

2.1.1 Contratações públicas

Contratos públicos podem ser entendidos como ajustes realizados entre a administração pública e particulares, para a execução de objetivos de interesse público, com

regras e condições estabelecidas pela própria administração. Administração pública é entendida como qualquer órgão ou entidade pública das esferas federal, municipal ou estadual [PIETRO, 2009]. Uma pessoa física ou jurídica de direito privado pode, de acordo com sua vontade, escolher quem ela deseja contratar. Não há legislação que estabeleça regras de como esse particular deva proceder nesse processo de escolha. Diferentemente de um particular, a Constituição Federal estabelece que os órgãos da administração pública devem seguir um procedimento que visa garantir a observância de critérios objetivos para a escolha do contratado, a transparência do processo de escolha, a isonomia, que é o tratamento igual entre os participantes, e a impessoalidade dessa escolha [Brasil, 1988].

A legislação vigente determina ainda que, como regra, a escolha do fornecedor ou contratado deve ser realizada por licitação pública. A licitação pode ser entendida como um procedimento administrativo pela qual as empresas interessadas concorrem entre si para serem escolhidas a fornecer um bem ou prestar serviço para um órgão público [TCU, 2010; Brasil, 1988, 1993, 2001]. As licitações públicas seguem regras bem rígidas. Essas regras têm origem na própria Constituição Federal de 1988, na legislação vigente, mais notadamente na Lei Federal nº 8.666, de 1993, que regulamenta as licitações, e na Lei Federal nº 10.520, que cria a modalidade do pregão para licitações, e no próprio edital da licitação [Brasil, 1988, 1993, 2001].

O edital é o instrumento pelo qual o órgão torna público seu interesse em contratar um serviço ou comprar um bem. Também é o edital que estabelece, entre outras coisas, todos os procedimentos, prazos, valores estimados e máximos, define o objeto a ser contratado ou comprado, e estabelece a modalidade da licitação.

As licitações públicas buscam sempre a proposta mais vantajosa para à administração pública e, para que esse objetivo seja atingido, é imprescindível a maior concorrência possível no processo licitatório. Maior número de licitantes, com propostas independentes, se traduz em contratações mais econômicas em benefício da sociedade [TCU, 2010]. Logo, para alcançar o princípio da eficiência previsto no texto constitucional, os gastos públicos precisam ter o menor custo possível, sem favorecer qualquer empresa, e com padrões de qualidade e eficiência bem estabelecidos em seu edital. Para que isso ocorra, é fundamental que as licitações tenham regras que facilitem a participação do maior número de concorrentes independentes [DPDC, 2008].

As contratações públicas envolvem grandes volumes de recursos econômicos. As

despesas de consumo do Governo respondem por cerca de 20% do Produto Interno Bruto (PIB) brasileiro nos últimos anos. Dados do Instituto Brasileiro de Geografia Estatística (IBGE), publicados no Relatório de Contas Nacionais referente ao quarto trimestre de 2016 quantificam esse percentual em aproximadamente 1,25 trilhão de reais [IBGE, 2017]. As contratações públicas se revestem de especial interesse para ações de controle pelo volume de recursos públicos envolvidos.

No contexto deste trabalho, é importante destacar que as licitações públicas são realizadas por órgãos públicos. Cada licitação pode conter um ou mais lotes. Cada lote corresponde a um produto ou serviço que é disputado individualmente pelas empresas para seu fornecimento. Desta maneira, em uma licitação em que haja vários lotes, uma empresa que seja muito competitiva em uma categoria de produto ou serviço pode disputar pelo o fornecimento de um único lote sem comprometer sua partição nos demais lotes. Tomemos como exemplo hipotético uma licitação que visa adquirir produtos alimentícios para a merenda escolar de um município. Uma licitação desse tipo é tipicamente dividida em lotes por categoria de produto, como hortifrúti, cereais, açougue e padaria. Assim, o órgão público consegue atrair empresas atacadistas que tipicamente se especializam no fornecimento de uma dessas categorias e, geralmente, conseguem ter um preço mais competitivo que empresas varejistas. Há outras divisões possíveis ainda, como por região geográfica, de maneira a atrair empresas fornecedoras que conseguem fornecedor um produto ou serviço a um preço competitivo, mas somente em uma região.

Uma série de estudos técnicos preliminares ao edital de licitação realizados pelo órgão licitante, determinando de que maneira a divisão por lotes deve ocorrer com vistas a aumentar a concorrência. A lei de licitações, Brasil [1993], deixa claro que, sendo tecnicamente viável e inexistindo prejuízo à economia de escala ou ao conjunto da contratação, as disputas licitatórias devem ser divididas em parcelas ou lotes, de forma a beneficiar o aumento da competitividade.

O pregão é o tipo de licitação preferido pelos órgãos públicos. De maneira geral ela funciona no modelo de leilão reverso. Diferente de um leilão tradicional em que o preço do bem adquirido vai aumentando até o maior lance vencedor, no pregão, o órgão público fixa para cada lote um valor máximo que está disposto a pagar. As empresas concorrentes fazem lances sucessivos e vão reduzindo o valor do lote. A empresa que, ao final desse processo, tiver a proposta com o menor valor é declarada vencedora [Brasil, 2001; DPDC, 2008].

A Figura 1 ilustra um modelo relacional possível para licitações públicas. O diagrama apresenta o relacionamento entre as entidades envolvidas no processo de licitações públicas, especificando também a relação dos sócios, contadores e funcionários com as empresas licitantes. Os dados armazenados nesta estrutura de dados serão modelados via grafos no Capítulo 4.

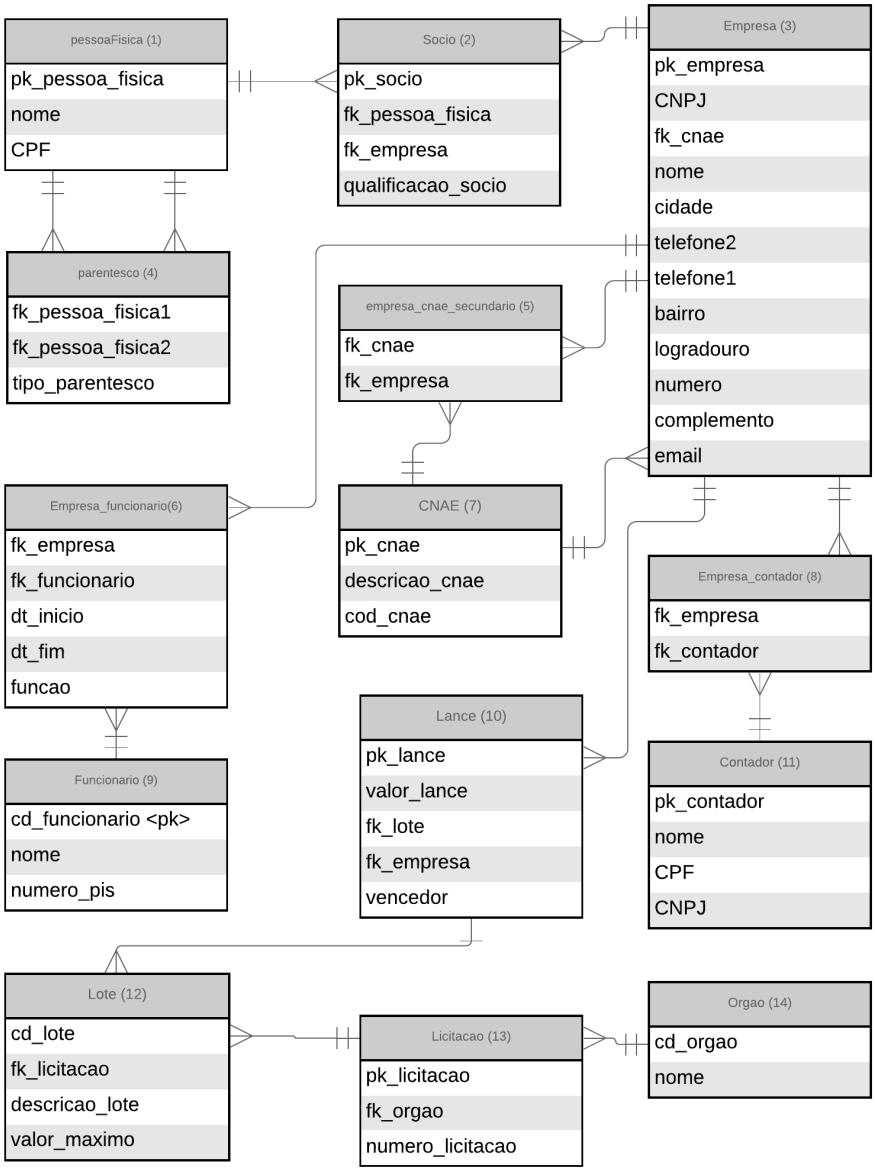


Figura 1 – Modelo Entidade Relaciona (MER) de um esquema de dados de licitações públicas

Neste modelo, os órgãos públicos organizam licitações públicas com o objetivo de adquirir bens ou produtos. As licitações públicas contêm um ou mais lotes de produtos ou serviços que os órgãos públicos desejam adquirir. As empresas participam das licitações

públicas com o objetivo de fornecer ao menos um lote da licitação. A participação da empresa em uma licitação se efetiva no momento em que a empresa dá um lance para ao menos um dos lotes contidos na licitação. Uma licitação só é considerada válida se tiver ao menos 3 (três) empresas participantes. Os lances são realizados de maneira decrescente no esquema de leilão reverso. A empresa que der o menor lance é declarada vencedora. As empresas possuem contadores que realizam a contabilidade das empresas. Funcionários trabalham para empresas. As empresas mantêm a sua CNAE primária e as suas CNAEs secundárias atualizadas. As empresas possuem um ou mais sócios. Esses sócios podem ter familiares como sócios em outras empresas. Os relacionamentos de participação societária, contadores, funcionários e familiares que representados neste modelo serão discutidos na Seção 2.1.2.

2.1.2 Irregularidade

As contratações públicas devem seguir uma série de normas já discutidas na seção 2.1.1. Qualquer infração ao rito da norma aplicável é uma irregularidade passível de sanção. As irregularidades de interesse deste trabalho são aquelas que resultam numa restrição à competitividade das licitações públicas e as de empresas que agem em conluio para fraudar a concorrência em uma licitação. A restrição à competitividade de licitação é quando as cláusulas contratuais, condições de fornecimento do bem ou condições ambientais do órgão licitante limitam, de maneira excessiva e desnecessária, as condições de participação de um número maior de empresas concorrentes [Sundfeld et al., 2018; TCU, 2010]. O conluio de empresas ocorre quando os proponentes, em vez de competirem, como seria de se esperar, conspiram secretamente para aumentar os preços.

A Constituição Federal não admite que as licitações contenham cláusulas restritivas à participação dos interessados. Veja o art. 37, XXI:

> *ressalvados os casos especificados na legislação, as obras, serviços, compras e alienações serão contratados mediante processo de licitação pública que assegure igualdade de condições a todos os concorrentes, com cláusulas que estabeleçam obrigações de pagamento, mantidas as condições efeti-*

> *vas da proposta, nos termos da lei, o qual somente permitirá exigências de qualificação técnica e econômica indispensáveis à garantia do cumprimento das obrigações [Brasil, 1988].*

Esta disposição é repetida no art. 3º, § 1º, I, da Lei n. 8.663/93:

> *É vedado aos agentes públicos admitir, prever, incluir ou tolerar, nos atos de convocação, cláusulas ou condições que comprometam, restrinjam ou frustrem o seu caráter competitivo, inclusive nos casos de sociedades cooperativas, e estabeleçam preferências ou distinções em razão da naturalidade, da sede ou domicílio dos licitantes ou de qualquer outra circunstância impertinente ou irrelevante para o objeto do contrato* [Brasil, 1993].

Então, por disposição constitucional e legal, as únicas exigências que a administração pode fazer dos interessados em licitar são aquelas indispensáveis ao cumprimento do contrato, sob pena de violação do princípio da competitividade. Toda licitação tem edital com cláusulas que restringem o objeto e o universo dos participantes, uma vez que a Administração necessita de um dado objeto, o que exclui os demais, semelhantes ou não, e de condições pessoais do futuro contratado que conduzam à alta probabilidade de que o contrato seja cumprido. As exigências não podem ir além do estritamente necessário à obtenção do objeto desejado pela administração pública [Sundfeld et al., 2018; TCU, 2010; Brasil, 1988, 1993].

Além de a competitividade no certame ser um princípio republicano, que busca a isonomia no tratamento entre as empresas, ela também tem reflexo direto no preço final da licitação. Quanto maior a competitividade, a tendência é de que o custo do bem ou serviço adquirido pela administração pública seja menor. Por outro lado, quanto menor a competitividade, maior o custo para aquisição do bem ou serviço [TCU, 2010].

A restrição de competitividade se reflete na participação das empresas licitantes. Observa-se um comportamento característico quando há restrições excessivas. Ao longo das licitações realizadas por órgãos contratantes, percebe-se que algumas empresas continuam participando das licitações normalmente enquanto outras participam algumas poucas vezes e não voltam a participar de outras licitações organizadas pelo mesmo órgão. É nesse comportamento característico que a metodologia da presente pesquisa tem seu foco [TCU, 2010; DPDC, 2008; Carvalho].

Uma característica marcante desse tipo de irregularidade é notada quando uma empresa favorecida se especializa em participar de licitações de um determinado órgão público. Independentemente do tipo do produto ou serviço relacionado no lote de licitação, a empresa favorecida sempre participa de licitações do órgão cooptado, mesmo que o produto ou serviço licitado esteja fora do seu ramo de atividade principal. Esse comportamento é, frequentemente, consequência de práticas anticompetitivas praticadas pelos atores das licitações públicas. De um lado está o órgão público cooptado, que criar regras anticompetitivas nos editais de licitação com o objetivo de favorecer uma empresa, e do outro lado, a empresa favorecida, que participa de todas as licitações organizadas pelo órgão cooptado [TCU, 2010; DPDC, 2008; Mendroni, 2009; Carvalho].

2.1.3 Classificação Nacional de Atividade Econômica

As definições a respeito da CNAE constantes desta seção foram retiradas de IBGE [2007].

A CNAE é a classificação oficialmente adotada pelo Sistema Estatístico Nacional na produção de estatísticas por tipo de atividade econômica, e pela Administração Pública, na identificação da atividade econômica em cadastros e registros de pessoa jurídica.

A CNAE tem como principal propósito ser uma classificação padronizada das atividades econômicas produtivas. Ela provê um conjunto de categorias para serem usadas na coleta e divulgação de estatísticas por tipo de atividade econômica. Essas categorias são definidas de acordo com a forma como o processo econômico está organizado nas unidades e como se quer que seja descrito nas estatísticas econômicas.

A CNAE, portanto, é usada para classificar as unidades de produção, de acordo com as atividades que desenvolvem, em categorias definidas como segmentos homogêneos, principalmente quanto à similaridade de funções produtivas (insumos, tecnologia, processos) e, em alguns casos, quanto às características dos bens e serviços ou, ainda, à finalidade de uso dos bens e serviços.

Toda empresa deve manter junto à Receita Federal o seu cadastro atualizado. Entre as informações cadastrais mantidas está a CNAE. O responsável pela empresa deve sempre atualizar duas informações relativas à CNAE: a CNAE principal, que repre-

senta a atividade econômica principal responsável pela atividade de maior receita, e a CNAE secundária, que são as atividades econômicas secundárias, as demais atividades exercidas na mesma unidade produtiva, além da atividade principal. Na Administração Pública, a CNAE é usada para a identificação da atividade econômica dos agentes produtivos nos cadastros e registros de pessoa jurídica. Em que pese a CNAE servir para classificar as atividades econômicas das unidades de produção, permite-se estabelecer uma correspondência entre a atividade econômica de origem e os produtos. A CNAE se aplica a todos as pessoas jurídicas no Brasil.

Entre as características da CNAE está a organização hierárquica, para possibilitar o uso para diferentes propósitos estatísticos. A CNAE é uma classificação estruturada de forma hierarquizada em cinco níveis, com 21 seções, 87 divisões, 285 grupos, 673 classes e 1301 subclasses.

A Tabela 1 resume a organização hierárquica da CNAE

Tabela 1 – Organização hierárquica do CNAE

Nível	**Número de grupamentos**	**Identificação**
1º	21	Código alfabético de 1 dígito
2º	87	Código numérico de 2 dígitos
3º	285	Código numérico de 3 dígitos
4º	673	Código numérico de 4 dígitos + DV
5º	1.301	Código numérico de 7 dígitos (incluindo o DV)

O modelo de codificação adotado na CNAE é misto, sendo formado de um código alfabético (uma letra), para indicar o primeiro nível de grupamento da classificação, a Seção, e de códigos numéricos para os demais níveis de agregação, divisão, grupo, classe e subclasse.

O sistema de codificação da CNAE é integrado somente a partir do segundo nível (divisão). Os dois primeiros dígitos numéricos representam a divisão. Os três primeiros dígitos numéricos representam o grupo. Os cinco primeiros dígitos numéricos representam a classe. E os sete primeiros dígitos numéricos representam a subclasse.

A CNAE será fundamental nesse trabalho para separar as licitações de acordo com a atividade econômica das empresas que participam da licitação. A relação entre as empresas e os órgãos públicos é muito influenciada pela atividade econômica das empresas. Essa distinção pela CNAE é importante para comparar somente empresas que exercem as mesmas atividades econômicas.

2.1.4 Tribunal de Contas

Não é objetivo desta seção detalhar todas as funções e competências dos tribunais de contas brasileiros, mas tão somente de fornecer uma visão geral sobre o funcionamento dessas entidades na fiscalização do dinheiro público.

A Constituição Federal de 1988 atribui aos Tribunais de Contas uma parcela do controle externo da administração pública. O controle pode ser classificado em interno, externo e social. O controle interno é realizado dentro da mesma estrutura hierárquica; o controle externo é exercido por um órgão para outro órgão; e o controle social é exercido pela sociedade na fiscalização dos atos da administração pública e seus gestores. O Tribunal de Contas é o órgão de controle externo mais operacional, atuando com relevante independência, o que lhe confere significativa isenção para apreciar as contas dos responsáveis por dinheiros, bens e valores públicos. O Tribunal de Contas da União é composto por 9 (nove) Ministros. Já os Tribunais estaduais são integrados por 7 (sete) Conselheiros [Brasil, 1988].

Na sua atuação, os Tribunais podem sustar atos administrativos reputados ilegais; exercer o controle de constitucionalidade de leis ou atos do Poder Público; adotar medidas cautelares, dentre outras providências. Portanto, os Tribunais de Contas atuam no sentido de garantir a moralidade dos atos e a probidade na gestão, fiscalizando a atuação dos administradores e dos agentes públicos em geral [Brasil, 1988; Sundfeld et al., 2018]. O Tribunal de Contas poderá fiscalizar e auditar todos os órgãos, entes da administração pública, pessoas físicas e pessoas jurídicas que de algum modo recebam, administrem ou gerenciem recursos públicos. A regra, portanto, é que o Tribunal de Contas, em tese, estará legitimado a atuar sempre que o caso envolver utilização, arrecadação, guarda, gerenciamento ou administração de bens, valores e recursos públicos [Sundfeld et al., 2018]

A Constituição Federal e as constituições de cada estado dispõem sobre a jurisdição dos Tribunais de Contas. A fiscalização dos recursos públicos federais é de responsabilidade do Tribunal de Contas da União. A fiscalização dos recursos públicos dos estados é de responsabilidade do respectivo Tribunal de Contas Estadual. A fiscalização dos recursos públicos municipais é de responsabilidade do respectivo Tribunal de Contas Estadual onde não houver Tribunal de Contas Municipal. O município do Rio de Janeiro

e o município de São Paulo contam, cada qual, com um Tribunal de Contas Municipal próprio, e as cidades dos estados do Pará, Bahia e Goiás contam com Tribunais de Contas dos municípios, que têm jurisdição sobre todos os municípios desses estados [Brasil, 1988; Sundfeld et al., 2018] .

A legislação prevê uma série de atribuições distintas aos Tribunais de Contas. Os Tribunais de Contas têm entre suas atribuições o dever de fiscalizar o gasto público. São inúmeros elementos passíveis de fiscalização dos Tribunais de Contas. editais de licitações públicas, quaisquer atos de contratação e desligamento de pessoal, a prestação das contas de governo, balancetes contábeis e contratos são alguns dos elementos sujeitos a fiscalização por parte do controle externo. Uma parcela representativa dos gastos públicos é oriunda das contratações públicas. Uma contratação pública tem vários pontos diferentes passíveis de fiscalização. Pontos ligados à economicidade, eficiência, eficácia e efetividade das contratações são alguns daqueles que podem ser analisados pelos Tribunais de Contas [TCE-RJ, 2010].

Os Tribunais de Contas responsabilizam os agentes públicos por irregularidades cometidas. Determinam, ainda, as formas de reparação dos eventuais danos ao erário identificados. Adicionalmente, produzem recomendações e determinam medidas preventivas para aumentar a eficácia da alocação dos recursos. Os Tribunais de Contas no Brasil têm ampla autonomia, com competência para elaborar o seu próprio programa de fiscalização [Speck, 2008].

Tomando como exemplo o TCE-RJ, o órgão é responsável por fiscalizar recursos públicos do Estado do Rio de Janeiro e de todos os municípios do estado do Rio de Janeiro, com exceção da capital, exatos 91 (noventa e um) municípios. Os seus 519 (quinhentos e dezenove) auditores são um quantitativo pequeno para fiscalizar em detalhes todos os aspectos relevantes de todas as contratações dos órgãos do estado e dos municípios [TCE-RJ, 2014]. Esses auditores são responsáveis por auditar todas as contratações dos órgãos jurisdicionados do TCE-RJ. Só o Estado do Rio de Janeiro, em 2015, gastou cerca de R$3,5 bilhões em contratações de bens e serviços para realizar as suas funções de Governo [TCE-RJ, 2016].

A capacidade dos órgãos de controle realizarem auditorias e fiscalizações é muitas vezes menor que o universo de contratos passíveis de fiscalização. Durante os anos de 2013 e 2018, os 743 (setecentos e quarenta e três) órgãos controlados pelo TCE-RJ celebraram uma média de pouco mais de 18.000 (dezoito mil) contratos por

ano. No ano de 2018, o TCE-RJ realizou 825 (oitocentos e vinte e cinco) auditorias com objetos e temas diversos, como auditorias para avaliação de execução de contratos, para verificação da registração contábil, para avaliação de políticas públicas, para avaliação do fluxo financeiro e para análise patrimonial. As auditorias sobre a execução de contratos são as que tipicamente abordam as irregularidades discutidas neste trabalho. Ainda no ano de 2018, o TCE-RJ realizou 49 (quarenta e nove) auditorias com o objetivo de apurar irregularidades em contratações públicas. Essas auditorias abrangeram um total de 43 (quarenta e três) órgãos do total de órgãos controlados pelo TCE-RJ. Isso corresponde a 5,78% de órgãos. O manual para seleção de objetos e ações de controle orienta a forma de escolha dos temas e o objetos. Essa escolha varia em função de diversos fatores, como capacidade técnica, capacidade operacional, materialidade, oportunidade, indicadores sociais, econômicos e ambientais. Portanto, o percentual de órgãos fiscalizados e os objetos de fiscalização escolhidos variam em função desses fatores[TCU, 2016; TCE-RJ, 2010, 2019; TCE-RJ, 2016].

Não há apuração da quantidade de contratos auditados, mas com auditorias realizadas em somente 43 (quarenta e três) dos 743 (setecentos e quarenta e três) órgãos possíveis, o número de fiscalizações é insuficiente para cobrir todas as contratações realizadas pelos órgãos controlados pelo TCE-RJ.

Em meio a um universo de contratações realizadas pelos órgãos públicos, os auditores selecionam uma pequena parcela dessas contratações para uma fiscalização pormenorizada com vistas à identificação de irregularidades. Para seleção dos trabalhos de auditoria, são usados critérios como risco (probabilidade de ocorrência de eventos futuros incertos com potencial para influenciar o alcance dos objetivos de uma organização), relevância (áreas consideradas estratégicas ou prioritárias nos instrumentos de planejamento governamental) e materialidade (importância relativa ou representatividade do valor ou do volume de recursos envolvidos) [TCE-RJ, 2010].

Achados de auditoria são fatos relevantes que representam desvios de normas ou procedimentos e cuja constatação decorre do processo de verificação e análise realizado pela auditoria. A aplicação de procedimentos e técnicas de auditoria visa à obtenção de evidências de auditoria. Constitui-se de investigações técnicas que permitem a formação fundamentada da opinião do auditor [TCE-RJ, 2010].

Técnicas de mineração de análise de dados, como as desenvolvidas no presente trabalho, auxiliam os auditores a selecionarem com maior acuidade os órgãos e as

contrações que devem sofrer auditoria. Após selecionarem os trabalhos a serem realizados, os auditores aplicam uma série de procedimentos a fim de identificar irregularidades ou inconformidades que vão ser relatadas em um relatório de auditoria [TCE-RJ, 2010; Balaniuk, 2010; TCU, 2016].

2.2- Teoria dos Grafos

Os conceitos básicos de grafos e a notação utilizada neste trabalho estão baseados no livro de Goldbarg and Goldbarg [2012].

Um grafo é uma estrutura de abstração bastante útil na representação e solução de diversos tipos de problemas. Matematicamente, um grafo formaliza relações de interdependência existentes entre os elementos de um conjunto. Um grafo pode ser entendido também como uma estrutura abstrata que representa um conjunto de elementos denominados vértices e suas relações de interdependência, que são denominadas arestas [Goldschmidt et al., 2015]. Ao longo deste trabalho, denotaremos um grafo por $G = (V, E, w)$, onde V é conjunto de vértices com cardinalidade n e E é o conjunto de arestas de G com cardinalidade m. Representa-se cada aresta conectando um vértice v_i ao um vértice v_j por (v_i, v_j) e um peso positivo w_{ij} é atribuído para cada aresta existente no grafo. Desta forma, os grafos que modelam a aplicação estudada nesta dissertação são grafos simples (sem laços e sem múltiplas arestas) e valorados nas arestas. A Figura 2 representa um grafo com 10 vértices e 10 arestas, de tal forma que $G = (V, E, w)$ onde $V = \{A, B, \ldots, J\}$ e $E = \{(A, B), (A, D), \ldots, (I, J)\}$. Arestas são adjacentes quando compartilham um vértice. Vértices são adjacentes se existe aresta os unindo Tomando como exemplo o grafo da Figura 2, as arestas (A, G) e (A, I) são adjacentes, por compartilhar o vértice A e os vértices G e I são adjacentes ao vértice A. Uma clique em um grafo não direcionado $G = (V, E)$ é um subconjunto de vértices C, tal que todos os vértices em C são adjacentes um ao outro. Ou seja, existe uma aresta conectando cada par de vértice. Isso se equivale a dizer que um subgrafo induzido de C é completo. O grau de um vértice é o número de arestas que incidem de um vértice. É denotado por $deg(v)$, onde v é o vértice analisado. Tomando como exemplo o grafo da Figura 2, $deg(A) = 2, deg(D) = 1$ e $deg(I) = 3$. A densidade de um grafo com n

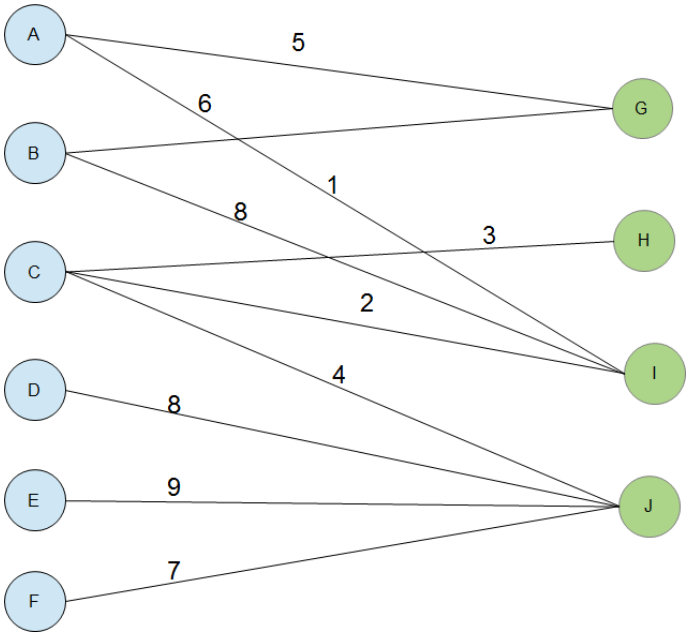


Figura 2 – Exemplo de grafo bipartido e valorado ($G = (V, E, w)$

vértices é a razão entre a quantidade de arestas do grafo e o total de arestas de um grafo completo com n vértices, ou seja, $d_G = 2m/n(n-1)$. Dependendo dessa proporção, um grafo é dito denso, quando possui muitas arestas para um determinada quantidade de vértices. Se o grafo possuir poucas arestas para uma determinada quantidade de vértices, o grafo é chamado de grafo esparso. A densidade do grafo Figura 2 é $d_G = \frac{2\times10}{10\times(10-1)}$. Um grafo é dito bipartido quando se pode particionar o conjunto de vértices do grafo em dois conjuntos V_o e V_e de tal forma que $V = V_o \cup V_e$ e $V_o \cap V_e = \emptyset$ e que todas as arestas do grafo estejam somente entre vértices de conjuntos diferentes. A Figura 2 ilustra um grafo bipartido. O conjunto $V_o = \{G, H, I, J\}$ e o conjunto $V_e = \{A, B, C, D, E, F\}$.

2.2.1 Mineração em grafos

Segundo Hand et al. [2001], a mineração de dados é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensíveis ao dono dos dados. As técnicas de mineração são usadas para extrair regras de conhecimento de um conjunto de dados brutos, utilizando o conhecimento obtido para realizar previsões, detectar

anomalias, reconhecer padrões, bem como diversos outros objetivos [Goldschmidt et al., 2015]. Neste processo de mineração, podemos extrair conhecimento não só de dados tabulares, mas também de dados estruturados de formas diferentes como a representação em grafos.

Qualquer relação M: N na terminologia do banco de dados pode ser representada como um grafo. Os dados de uma ampla variedade de disciplinas podem ser modelados em grafos. As redes de computadores, em computadores podem representar vértices e os links em arestas. As redes sociais consistem em indivíduos e suas interconexões, que podem ser relações comerciais, parentesco ou confiança, etc. As redes de interação proteica vinculam proteínas que devem trabalhar juntas para desempenhar alguma função biológica específica. As redes alimentares ecológicas vinculam espécies a relações predador-presa. Licitações públicas vinculando empresas privadas e órgãos licitantes, como é o caso deste trabalho. Nestes e em muitos outros campos, a abordagem por grafos é apropriada [Chakrabarti and Faloutsos, 2006].

O uso de grafos para representar os dados oferece suporte à visualização direta, aumenta a compreensibilidade do conhecimento e permite o uso de algoritmos aplicáveis exclusivamente a grafos como os algoritmos de detecção de comunidades que serão vistos na Seção 2.2.2. Portanto, a mineração de grafos é uma das abordagens mais promissoras para extrair conhecimento de dados relacionais Cook and Holder [2006].

2.2.2 Detecção de comunidades

O objetivo da detecção de comunidades é identificar subconjuntos de vértices em um grafo G que estão mais fortemente relacionados entre si do que com os demais vértices. Esse conjunto de vértices é denominado comunidade ou cluster [Goldschmidt et al., 2015]. Essa maior concentração de arestas correlacionando um determinado grupo de vértices em detrimento de outros vértices é uma características de redes reais. Essa falta de homogeneidade tem potencial para revelar propriedades comuns aos vértices pertencentes à mesma comunidade [Fortunato, 2010]. Família, vizinhos, círculos de amizade, trabalho e grupo de empresas são exemplos que se formam de comunidades dentro de outras redes de relacionamentos já existentes [Sun and Sun, 2017], [Fortunato,

2010]. A Figura 3 ilustra esse tipo de grafo, onde as áreas em cinza indicam comunidades detectadas.

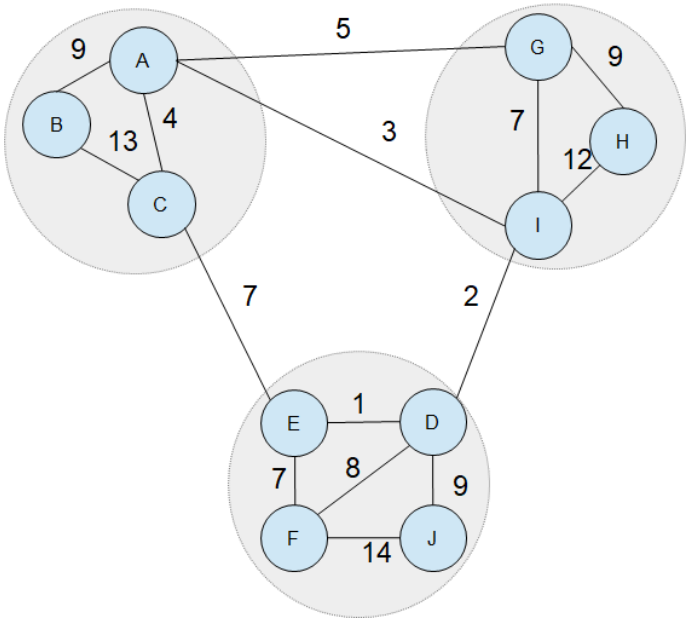


Figura 3 – Exemplo de comunidades detectadas detectadas num grafo

No presente trabalho, a detecção de comunidades a partir dos algoritmos de Girvan-Newman (GN) e de Clauset, Newman e Girvan (CNM) são as técnicas a serem aplicadas com o objetivo de identificar a concentração de relacionamentos e interações entre os órgãos públicos, empresas contratadas e seus sócios e empregados. Uma comunidade detectada em meio a um grafo que represente essas relações de compra e contratação dos órgãos públicos com as empresas fornecedoras é um indicador importante da presença da atuação de um cartel de empresas. Utilizamos ainda a medida de modularidade, proposta por [Newman, 2010], para avaliar a qualidade de cada agrupamento obtido.

2.2.3 Agrupamento: o método de Girvan-Newman (GN)

Um método bem difundido para detecção de comunidades é o método Girván-Newman (GN) de Girvan and Newman [2002], que propõe um método divisivo para detecção de comunidades em um grafo G. Ou seja, a solução começa com uma partição

com todos os vértices em uma única comunidade e, progressivamente, o grafo é dividido em k_i comunidades. Nesse método, as arestas são removidas progressivamente. As arestas mais acionadas para se caminhar de um vértice a outro no grafo G, ou seja, as arestas mais importantes para a conectividade do grafo G é o critério de ordem para a sua remoção. E, a cada remoção de aresta, deve se verificar se os subgrafos formados respondem ao problema modelado ou se as comunidades encontradas a cada iteração satisfazem a detecção esperada.

O método de GN usa o valor de intermediação como critério para remoção das arestas. A intermediação de uma aresta (v_i, v_j) é uma medida de centralidade. Ela quantifica o número de vezes em que uma aresta serve de caminho entre os vértices do grafo [Girvan and Newman, 2002]. Uma aresta com alto valor de intermediação representa um dos elos de ligações entre duas partes de uma rede, cuja remoção pode afetar a comunicação entre vários pares de nós pelos caminhos mais curtos entre eles [Lu and Zhang, 2013]. O valor de intermediação de uma aresta é dada pela seguinte expressão:

$$c_B(e) = \sum_{s,t \in V} \frac{\sigma(s,t|e)}{\sigma(s,t)} \tag{1}$$

onde V é o conjunto de nós do grafo, $\sigma(s,t)$ é o número de caminhos mais curtos entre s e t no grafo, e $\sigma(s,t|e)$ é o número de caminhos mais curtos que conectam s e t e passam pela aresta e. Tomemos como exemplo o grafo da Figura 4. Note que cada uma das arestas tem um valor de intermediação associado que denota um percentual de vezes em que ela serve de caminho mais curto entre dois vértices distintos do grafo. A Tabela 2 apresenta a memória de cálculo para computar o valor de intermediação de cada aresta. Na primeira coluna, tem-se o vértice de fonte e de destino. Na segunda coluna, tem se o caminho mais curto baseado no peso de cada aresta. Cada coluna na região hachurada representa uma aresta e em que momento ela é computada dado o caminho escolhido. A última linha apresenta o valor de intermediação de cada aresta.

Uma aresta com o valor de intermediação alto representa um elo em forma de ponte entre duas partes de uma rede, cuja remoção pode afetar a comunicação entre muitos pares de nós.

No exemplo, a remoção da aresta (D, B), que tem o maior valor de intermediação, resulta em uma partição da rede em duas comunidades densamente conectadas: as comunidades $\{A, B, C\}$ e $\{D, E, F\}$. O Algoritmo 1 apresenta um pseudo código para o

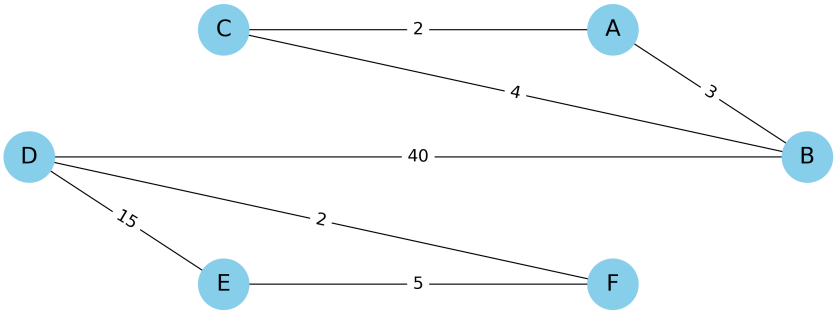


Figura 4 – Exemplo de grafo

Tabela 2 – Memória de cálculo para computar o valor de intermediação das arestas do grafo de exemplo da Figura 4

			Número de vezes em que a aresta (v_i, v_j) é usada como caminho						
		Caminho mais curto	(A,B)	(A,C)	(C,B)	(B,D)	(D,E)	(F,E)	(D,F)
1	A-B	(A,B)	1						
2	A-C	(A, C)		1					
3	A-D	(A,B), (B,D)	1			1			
4	A-E	(A,B), (B,D), (D,F), (F,E)	1			1		1	1
5	A-F	(A,B), (B,D), (D,F)	1			1			1
6	B-C	(B, C)			1				
7	B-D	(B, D)				1			
8	B-E	(B, D), (D, F), (F,E)				1		1	1
9	B-F	(B, D), (D, F)				1			1
10	C-D	(C,B), (B, D)			1	1			
11	C-E	(C,B),(B, D), (D, F), (F,E)			1	1		1	1
12	C-F	(C,B) ,(B, D), (D, F)			1	1			1
13	D-E	(D,F),(F,E)						1	1
14	D-F	(D, F)							1
15	E-F	(E, F)						1	
Intermediação da aresta v_i, v_j			$\frac{4}{15}$	$\frac{1}{15}$	$\frac{4}{15}$	$\frac{9}{15}$	$\frac{0}{15}$	$\frac{5}{15}$	$\frac{8}{15}$

método GN. O Algoritmo 1 recebe como entrada o grafo G e enquanto houver arestas no grafo G, a aresta de maior valor de intermediação é identificada e removida grafo G. Os conjuntos de vértices que ficarem desconexos são as comunidades C_{K_i} detectadas para a partição p_i. Cada configuração possível de agregação dos vértices em comunidades que o método retorna é uma partição. Uma partição p_i é um conjunto de comunidades C_{K_i}.

Note que, a cada iteração, o valor de intermediação para o grafo G é recalculado. Essa etapa é necessária, tendo em vista que as arestas de um caminho secundário podem ter seu valor de intermediação aumentado após a remoção de uma aresta de uma caminho importante. Dada a necessidade de se calcular o valor de intermediação das

arestas do grafo a cada iteração, a complexidade computacional desse algoritmo é de $O(|E|^2.|V|)$. Essa é a grande desvantagem desse método.

Algoritmo 1 – Girvan e Newman algoritmo (G)

Output: $p = [\,]$ - vetor de partições

```
1: i = 0
2: while thereIsEdgesIn(G) do
3:    vivjH ← highestEdgeBetweenness(G)
4:    removeEdge(G, vivjH)
5:    C_{k_i} ← disconnectedVertexSets(G)
6:    if C_{k_i} ≠ ∅ then
7:       if p = ∅ then
8:          p[i] ← C_{k_i}
9:       else
10:         if C_{k_i} ≠ p[i − 1] then
11:            p[i] ← C_{k_i}
12:         end if
13:      end if
14:   end if
15:   i ← i + 1
16: end while
17: return p
```

2.2.4 Qualidade de um agrupamento: método da modularidade

Não há, no método GN, um critério de parada. Pode-se remover aresta por aresta, progressivamente, até a rede não conter mais arestas. Portanto, o método GN retorna como resposta p_i partições como resposta para o grafo G. Uma forma de avaliar a qualidade de cada partição e selecionar uma delas é o valor de modularidade da partição, denotado por $Q(G, p_i)$ [Newman, 2010].

A modularidade é a diferença entre a soma dos pesos das arestas existentes e a esperada dentro das comunidades. Assim, uma partição com alta modularidade indica que a densidade de arestas dentro das comunidades é maior que o esperado ao acaso, indicando uma boa partição da rede [Newman, 2010]. Cabe destacar que há abordagens diferentes para computar o valor da modularidade. Em nosso trabalho utilizamos a abordagem de Newman [2010][p.224] para calcular a modularidade. Denotamos a modularidade pela Expressão 2.

$$Q(G, p_i) = \frac{1}{2m} \sum_{ij \in E} \Big(A_{ij} - \frac{k_i k_j}{2m} \Big) \delta(c_i, c_j), \tag{2}$$

onde m é a soma dos pesos das arestas; A_{ij} é a entrada (i, j) da matriz de adjacência de G; k_i é a soma dos pesos das arestas adjacentes ao vértice v_i; k_j é a soma dos pesos das arestas adjacentes ao vértice v_j e $\delta(c_i, c_j)$ é uma função que retorna 1, se v_i e v_j pertencerem a mesma comunidade e zero, caso contrário.

2.2.5 Agrupamento: o método de Clauset, Newman e Girvan (CNM)

O método CNM proposto por Clauset et al. [2004] é um método aglomerativo e utiliza estruturas de dados diferentes de GN para alcançar um tempo de execução melhor. Clauset et al. [2004] defendem que para a maior parte dos grafos do mundo real a complexidade de seu método é de aproximadamente $O(nlog^2 n)$, onde n é o número de vértices do grafo, o que permitiria a aplicação em grafos da ordem de milhões de vértices. A estratégia básica adotada pelo método é unir progressivamente os pares de vértices e comunidades que incrementem a modularidade da partição. Se comparado ao método GN, o método proposto por Clauset et al. [2004] tem importantes diferenças, a saber: 1) é um método aglomerativo, ou seja, o método inicializa com a partição solução onde cada vértice pertence a sua própria comunidade e, a cada iteração do método, as comunidades são fundidas; 2) a cada iteração, o método procura aumentar a modularidade da partição solução; 3) ele tem um critério de parada claro, que é quando o método não consegue aumentar a modularidade da partição solução pela fusão de comunidades; 4) faz uso de outras estruturas de dados, como árvores binárias e filas de prioridade, a fim de otimizar o consumo de memória.

A operação do algoritmo envolve encontrar as alterações em Q a cada fusão de pares de comunidades, escolhendo a maior delas e realizando a fusão correspondente. Para tanto, o método CNM centra a tomada de decisão para fusão das comunidades em duas estruturas de dados, uma matriz esparsa, contendo ΔQ_{ij} e um vetor a_i

ΔQ_{ij} é dada pela seguinte equação:

$$\Delta Q_{ij} = \begin{cases} \frac{1}{2m} - \frac{k_i k_j}{(2m)^2}, & \text{se } v_i \text{ e } v_j \text{ são adjacentes,} \\ 0, & \text{caso contrário.} \end{cases} \tag{3}$$

ΔQ_{ij} pode ser considerado como o incremento na modularidade que teríamos ao fundir os pares de comunidades em uma só comunidade.
a_i é dado pela seguinte equação:

$$a_i = \frac{k_i}{2m} \tag{4}$$

O método CNM pode ser definido nos seguintes passos:

1. Calcula os valores iniciais de ΔQ_{ij} e a_i de acordo com as equações 3 e 4;
2. As comunidades correspondentes ao maior valor da matriz ΔQ_{ij} e a_i são fundidas;
3. A matriz ΔQ_{ij} e a_i e o vetor a_i são atualizados;
4. Os passos 2 e 3 são repetidos até haver somente uma comunidade ou não ser mais possível fundir comunidades que aumentem a modularidade.

O passo 3 de atualização da matriz ΔQ_{ij} é realizado pelas seguintes regras:

$$\Delta Q_{jk} = \begin{cases} \Delta Q_{ik} + \Delta Q_{jk} \text{ se a comunidade } k \text{ está conectado a } i \text{ e } j, \\ \Delta Q_{ik} - 2a_j a_k \text{ se a comunidade } k \text{ está conectado somente a } i, \\ \Delta Q_{jk} - 2a_i a_k \text{ se a comunidade } k \text{ está conectado somente a } j, \end{cases} \tag{5}$$

Clauset et al. [2004] propõem que para a implementação do método CNM, a estrutura de dados para cada linha da matriz ΔQ_{ij} seja uma árvore binária balanceada. Dessa maneira, a árvore binária permite que cada elemento possa ser encontrado e inserido em tempo de $O(logn)$. Essa melhoria resulta em uma considerável economia de memória e tempo. Além da árvore binária para cada linha da matriz, uma $heapH$ contendo o elemento de maior valor para cada linha da matriz ΔQ_{ij} auxilia na função de recuperar o elemento de maior valor em um tempo constante.

2.3- Medida de entropia

Há uma diversidade de definições e aplicações do conceito de entropia na ciência. Mais recentemente, este conceito de entropia tem sido utilizado como medida de identificação de anomalias em grafos. Em Liu et al. [2016], os autores utilizaram a definição de entropia para a identificação de conluios no que diz respeito a médicos com prescrições de narcóticos e as respectivas farmácias que venderam esses produtos. Inspirados neste trabalho, utilizamos a entropia para a identificação de licitações com restrição à competitividade. Neste contexto, a entropia está associada à distribuição dos graus dos vértices levando-se em consideração o peso de cada aresta. Dado um vértice v_i, definimos $N(v_i)$ como o conjunto dos vizinhos de v_i. O peso de uma aresta $e_{ij} = (v_i, v_j)$ é denotado por w_{ij} e o grau de v_i é dado por $\deg(v_i) = \sum_{k \in N(v_i)} w_{ik}$. A partir destes parâmetros, a entropia de um vértice v_i é definida como:

$$H_{v_i} = \frac{1}{log\,(deg(v_i))} \sum_{v_k \in N(v_i)} p_k log \frac{1}{p_k}, \tag{6}$$

onde $N(v_i)$ é o conjunto de vizinhos de v_i e

$$p_k = \frac{w_{i,k}}{\sum_{k \in N(v_i)} w_{i,k}}$$

é o percentual de relacionamento do vértice v_i com o vizinho v_k sob o total de relacionamentos do vértice v_i O resultado é a entropia empírica, medindo a dispersão de relacionamentos do vértice v_i entre seus vizinhos $N(v_i)$. A entropia é dividida ainda por $log(\deg(v_i))$ para normalizar para o intervalo $[0, 1]$. Quanto mais uniforme a distribuição dos relacionamentos de um vértice v_i, mais próximo de 1 será a razão da entropia. Se o vértice v_i, que representa um órgão licitante tiver como participantes de suas licitações sempre as mesmas empresas, a razão de entropia é 1. Se, ao contrário, v_i tiver como participantes de suas licitações mais uma empresa do que outras, a dispersão será muito distorcida, resultando em uma taxa de entropia mais próxima de 0. Uma entropia baixa reflete o comportamento característico em licitações com restrição a competitividade de certame, conforme apresentado na Seção 2.1.2. Desta forma, na aplicação que abordaremos aqui neste trabalho, espera-se que a participação das empresas concorrentes nas

licitações organizadas pelo mesmo órgão licitante para a mesma atividade econômica seja homogênea.

2.4- Teste de Aderência de Kolmogorov-Smirnov

O teste de aderência de Kolmogorov-Smirnov (K-S) será utilizado para verificar a aderência das distribuições estatísticas dos dados reais de contratações públicas. Essa informação da distribuição estatísticas será importante para simular novas redes com dados sintéticos e com as mesmas características das redes reais. As explicações referentes a esta seção foram retiradas de Morettin and Bussab [2017]; Massey [1951].

O teste de K-S avaliará a aderência de uma amostra de dados a uma dada distribuição hipotética. O teste verifica a hipótese de uma população ter uma distribuição teórica específica. Assumir que determinado grupo de dados se distribui conforme um modelo nos permite realizar estimativas sem precisar da totalidade das informações.

O teste de K-S pode ser utilizado para avaliar as seguintes hipóteses:

Hipótese nula (H_0): A distribuição testada pode ser utilizada para prever o comportamento dos dados observados.

Hipótese alternativa (H_1): A distribuição testada **não** pode ser utilizada para prever o comportamento dos dados observados.

Este teste observa a máxima diferença absoluta entre a função de distribuição acumulada esperada para H_0 e a distribuição observada dos dados. Como critério, comparamos esta diferença com um valor crítico, para um dado nível de significância. Se o valor máximo da diferença absoluta não for superior ao valor crítico para o nível de significância testado, a hipótese nula (H_0) não é rejeitada.

Considere uma amostra aleatória simples X_1, X_2, X_n de uma população com função de distribuição acumulada (CDF) contínua desconhecida. Designaremos por $f(x)$ a função de densidade, e por F_X a função de distribuição acumulada (CDF) de X. Estimar f_X é equivalente a estimar F_X. O objetivo é testar se a amostra observada veio de uma distribuição de probabilidades esperada. $H_0 = F(x) = F_0(x)$, para todo (x). Considere a função de distribuição empírica $F_e(x)$, como um estimador de $F(x)$, para todo valor x real. Considere a diferença absoluta como sendo $D_\alpha(N) = \max_x |F(x_i) - F_e(x_i)|$. Se $F_e(x)$ for

um bom estimador de $F(x)$, então $D_\alpha(N)$ será um valor baixo. Onde N é o tamanho da amostra e α é o nível de significância requerido.

Há várias maneiras de medir a "distância"entre F_X e $F_e(x)$. o teste K-S propõe uma estatística para o teste, obtida tomando o máximo dos valores absolutos das diferenças $|F(x_i) - F_e(x_i)|$. No caso,

$D = \max_{1 \leq i \leq n} |F(x_i) - F_e(x_i)|$.

O valor encontrado deve ser comparado com o valor crítico, para um dado nível de significância e número de elementos da amostra. Se D é maior que o valor crítico $D_\alpha(N)$, rejeita-se H_0 dos dados. Caso contrário, não se rejeita a hipótese nula. Tipicamente, se utiliza 5% como nível de significância ou $\alpha = 0.05$. A tabela 3 apresenta os valores críticos para o teste K-S

Tabela 3 – Valores críticos, $D_\alpha(N)$ (fonte: Massey [1951])

	Significância (α)		
Tamanho da amostra (N)	**0,05**	**0,02**	**0,01**
5	0,563	0,627	0,669
10	0,409	0,457	0,589
15	0,338	0,377	0,404
20	0,294	0,329	0,352
25	0,264	0,295	0,317
30	0,242	0,270	0,290
35	0,224	0,254	0,273
Valores maiores que 35	$\frac{1,358}{\sqrt{N}}$	$\frac{1,517}{\sqrt{N}}$	$\frac{1,628}{\sqrt{N}}$

2.5- Boxplot

O boxplot é um gráfico que possibilita representar a distribuição de um conjunto de dados. Essa ferramenta será importante para identificar outliers. Os auditores podem se concentrar em uma análise preliminar nas contratações de órgãos cujo valor de entropia esteja fora do intervalo entre o primeiro e segundo quartil.

Os parâmetros descritivos para o boxplot são: a mediana (Q_2), o quartil inferior (Q_1), o quartil superior (Q_3) e do intervalo interquartil $(IQ_R = Q_3 - Q_1)$. A figura 5 mostra um exemplo

A linha central da caixa marca a mediana do conjunto de dados. A parte inferior

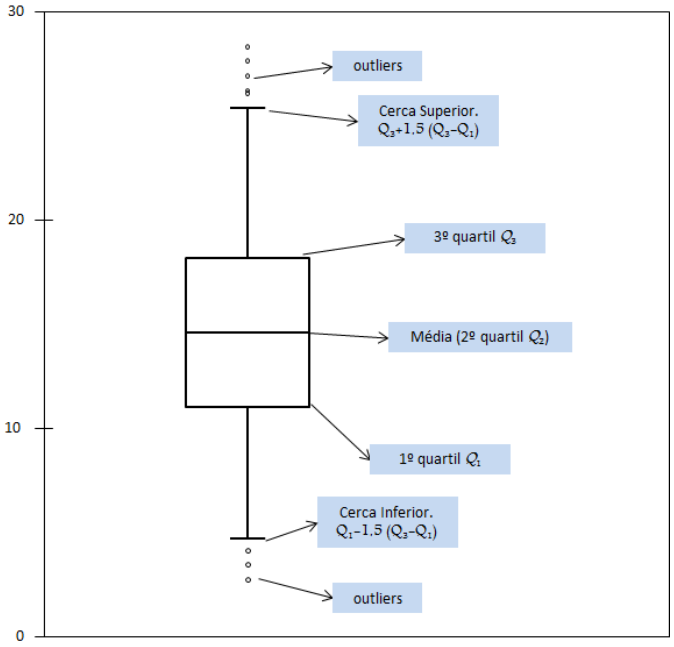


Figura 5 – Exemplo de boxplot

da caixa é delimitada pelo quartil inferior (Q1) e a parte superior pelo quartil superior (Q3). As hastes inferiores e superiores se estendem, respectivamente, do quartil inferior até o menor valor não inferior a $Q_1 - 1.5 \times IQ_R$ e do quartil superior até o maior valor não superior a $Q_3 + 1.5 \times IQ_R$. Os valores inferiores a $Q_1 - 1,5 \times IQ_R$ e superiores a $Q_3 + 1.5 \times IQ_R$ são representados individualmente no gráfico sendo estes valores caracterizados como outliers.

As quantidades $Q_1 - 1.5 \times IQ_R$ e $Q_3 + 1.5 \times IQ_R$ delimitam, respectivamente, as cercas inferior e superior e constituem limites para além dos quais, como visto, os dados passam a ser considerados outliers.

O boxplot permite avaliar a simetria dos dados, sua dispersão e a existência ou não de outliers nos mesmos, sendo especialmente adequado para a comparação de dois ou mais conjuntos de dados correspondentes às categorias de uma variável qualitativa.

3- Trabalhos correlatos

Conforme introduzido no capítulo 1, devido à importância do tema, a identificação de fraudes em licitações públicas tem sido abordada em diversos trabalhos. Foram realizadas buscas em duas bases de maneira a estabelecer um mapa sistemático para a avaliação da literatura, o Science Direct e o IEEE. Foram realizadas pesquisas nas bases de busca a fim de identificar trabalhos cujo objetivo fosse a identificação de fraudes em aquisições governamentais.

Realizou-se uma pesquisa na base de busca do Science Direct com os seguintes parâmetros: a string “(government) AND (procuring OR acquisition OR procurement) AND (predict OR detect OR detection OR forecast OR ”data mining”)” foi utilizada no campo “Title, abstract or keywords” em 10 de outubro de 2018. Foram obtidas 83 referências. Realizou-se uma pesquisa na base de busca do IEEE, em 12 de outubro de 2018, com a string “(government) AND (procuring OR acquisition OR procurement) AND (predict OR detect OR detection OR forecast OR ”data mining”)”. Foram obtidas 252 referências.

Foram obtidas ao todo 335 referências. Foram descartados artigos que não tratavam de detecção de fraudes ou anomalias, ou artigos que não tinham uma abordagem afeta à mineração de dados. Após esse descarte, foram selecionados 6 trabalhos.

Ademais, realizou-se uma busca nos seminários do evento Brasil Digital. Esse evento tem por objetivo promover o compartilhamento de experiências e boas práticas relacionadas ao uso de técnicas de análise e mineração de dados como instrumento para melhoria da gestão e do controle de entidades e políticas públicas. Esse evento é organizado por órgãos responsáveis pela fiscalização dos gastos públicos de órgãos da administração pública brasileira. Desse evento, foram selecionados dois trabalhos para uma análise mais detida.

A técnica de eyeballing foi aplicada nesses 8 trabalhos selecionados, de onde se originaram mais dois trabalhos relacionados à presente pesquisa, que têm seu foco na detecção de fraudes em contratações públicas por meio de técnicas afetas à mineração de dados. A seguir, uma breve descrição desses 10 trabalhos é apresentada.

O trabalho de Padhi and Mohapatra [2011] tem como objetivo a detecção de cartéis e conluios de empresas em compras públicas do governo da Índia. A abordagem

da pesquisa está baseada em testes estatísticos sobre as proporções do preço de oferta estimado e do preço da oferta vencedora para detectar conluio. A abordagem precisa dividir as proporções em dois grupos significativamente diferentes. O grupamento com maior média e mediana, maior variância, assimetria negativa e auto-correlações significativas correspondem à licitação com conluio, enquanto o outro agrupamento corresponde à licitação competitiva.

O trabalho de Carvalho et al. [2014] tem o objetivo de identificar fraudes nas contratações públicas em órgãos da administração pública federal brasileira, mais especificamente dois tipos de fraudes: (1) se duas empresas que participam do processo de aquisição têm proprietários que são parceiros; ou (2) se uma única contratação fora fracionada em vários pequenos contratos, com valores abaixo R$ 8.000,00 (Oito Mil Reais), a fim de se evitar um processo licitatório. Os pesquisadores têm uma abordagem orientada pelo Cross Industry Standard Process for Data Mining (CRISP-DM) para orientar o processo de mineração de dados. Para desenvolver o modelo preditivo, a pesquisa usa o classificador Naive-Bayes e Redes Bayesianas. O modelo gerado foi capaz de classificar corretamente todas as compras divididas e uma área realmente alta do ROC (0,999).

O trabalho de Domingos et al. [2016] tem como objetivo detectar anomalias de contratações de bens e serviços de tecnologia da informação pelo governo federal brasileiro. Assim como na abordagem da pesquisa de Carvalho et al. [2014], os pesquisadores trabalham como CRISP-DM para orientar o processo de mineração de dados. A abordagem escolhida é por um algoritmo não supervisionado de aprendizagem profunda para gerar um modelo preditivo. O algoritmo escolhido foi o autoencoder. A partir de 18 atributos relativos às contratações de bens e serviços de tecnologia da informação foi construído um modelo para identificação de contratações anômalas. Com o auxílio desse modelo, é possível identificar contratações anômalas no conjunto de dados e, em seguida, avaliar seus atributos para verificar se o evento merece ser priorizado para investigação adicional pelo órgão de fiscalização responsável no âmbito do governo federal brasileiro.

O trabalho de Ralha and Silva [2012] tem como objetivo a detecção de cartéis econômicos de empresas em licitações públicas realizadas pelo governo federal brasileiro. Para tanto, os autores utilizam métodos baseados em regras de agrupamento e associação, e uma abordagem multiagente para descobrir as estratégias dinâmicas de empresas envolvidas na formação de cartéis.

O trabalho de Erven et al. [2017] tem o propósito de encontrar fraudes em licitações realizadas pelo governo brasileiro. O tipo de fraude aqui pesquisada é a fraude por relacionamentos de sócios ou parentescos destes sócios entre empresas participantes de um mesmo processo de compra. Os autores comparam a pesquisa manual, a pesquisa pelo banco de dados NoSQL e a pesquisa por banco de dados relacional. O propósito aqui é verificar qual metodologia é mais indicada ao uso pelo órgão responsável pela fiscalização dessas licitações. São comparadas questões de performance e acessibilidade no uso.

A pesquisa de Arief et al. [2016] tem como objetivo a detecção de possíveis fraudes que ocorrem no processo de aquisição através do sistema de acompanhamento da Indonésia. Para tanto, foram utilizados técnicas de mineração de dados baseadas em aprendizado supervisionado. Devido à ausência de rótulos nas bases de dados disponíveis, os autores se utilizaram de informações não estruturadas, como decisões judiciais e comentários públicos, a fim de rotular dados da base de dados utilizadas no processo de mineração de dados. Foram utilizados os métodos Naive Bayes, rede bayesiana, redes neurais e árvore de decisão. A pesquisa conclui que o método Naive Bayes apresenta os melhores resultados para o problema pesquisado.

O trabalho de Davydenko et al. [2017] tem seu foco na descoberta de esquemas de fraude em aquisições publicas realizadas pelos governos. A abordagem dos pesquisadores está centrada em técnicas de mineração de texto, análise de agrupamento e de mineração de dados. Estas técnicas são utilizadas para revelar grupos de usuários propensos a cometer fraudes financeiras na esfera das transações de contratos públicos. O objetivo da pesquisa foi testar e usar técnicas de análise de agrupamento em relação a contratos com o governo, a fim de melhorar o reconhecimento de acordos de má fé. As principais ferramentas utilizadas foram o Oracle Data Mining e o IBM i2. Neste trabalho, entre as análises realizadas, se destacam as seguintes: centralidade de intermediação; centralidade de proximidade; centralidade do grau; centralidade do autovetor.

O trabalho de Fraga et al. [2017] tem como objetivo a detecção de casos suspeitos de fraudes em licitações realizadas por municípios do estado da Paraíba, mais especificamente identificar a presença de empresas em diferentes grupos com alta propensão de vitória de participantes; mapear grupos de empresas com indícios de simulação de concorrência; investigar indícios de concentração regional na atuação de grupos de empresas (segmentação de mercado ou direcionamento de licitações); ranquear empresas conforme

recorrências em grupos suspeitos de conluio, levando em conta a tendenciosidade de vitória, a presença de falsos concorrentes e a concentração regional. A abordagem da pesquisa se utiliza de métodos afetos a regras de associação com padrões de estratégias cooperativas, especialmente pela mensuração da probabilidade de vitória de grupos de empresas com participação em conjunto em número muito superior ao da média dos demais concorrentes. Foi utilizado o algoritmo Apriori com este propósito. A pesquisa detectou diversos possíveis cartéis econômicos de empresas atuando.

O trabalho de Souza and Pereira [2009] propõe a identificação de comportamentos suspeitos nos sistemas brasileiros de compras governamentais eletrônicas. O objetivo em específico é identificar conluios entre funcionários públicos e empresas participantes das licitações. A abordagem da pesquisa é por algoritmos afetos a regras de associação. Tal trabalho apontou para a necessidade de se utilizar mais informações dos fornecedores bem como se testar novas hipóteses.

O trabalho Melo and Ferreira [2016] busca identificar cartéis de empresas. A abordagem dos autores utiliza técnicas afetas a grafos e análise de redes sociais, mais especificamente, Pagerank e detecção de agrupamentos.

O trabalho de Carvalho et al. [2013] tem como objetivo a detecção de possíveis fraudes em aquisições com recursos do governo federal. As agências de fiscalização enfrentam o problema de classificação dos dados disponíveis em conhecimento útil. O modelo desenvolvido na pesquisa foi capaz, para dois casos, de transformar as informações disponíveis e classificá-las de acordo com os cenários e critérios. Para tanto, os pesquisadores usam Probabilistic Ontology Web Language para projetar e testar um modelo que realize a fusão de informações disponíveis aos órgãos de fiscalização e a sua classificação com base em uma modelo de probabilidades. Para projetar este modelo, uma ferramenta recentemente desenvolvida para criar ontologias foi usada com suporte de especialistas em Probabilistic Ontology Web Language.

A Tabela 4 compara as abordagens relacionadas anteriormente quanto às abordagens utilizadas na presente pesquisa. Diferentemente das pesquisas apresentadas nesse capítulo, a presente pesquisa é a única que tem uma abordagem centrada na mineração por grafos e teoria da informação para a detecção de indícios de fraudes em contratações públicas. Ademais, esta pesquisa é a única que tem seu foco mais especificamente na identificação de possíveis casos de restrição à competitividade em contratação públicas.

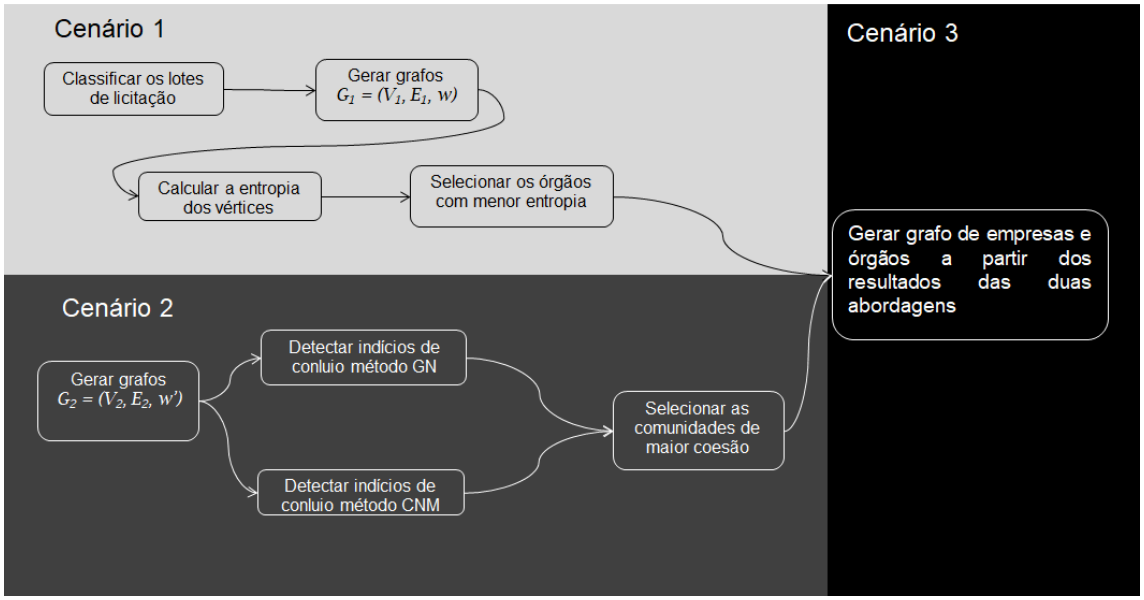
Tabela 4 – Tabela comparativa da presente pesquisa com trabalhos correlatos.

	Modelo via teoria de grafos	Identificação de restrição a competitividade	Identificação de conluio	Uso de regras de associação	Uso de detecção de comunidades	Uso de teoria da informação
Carvalho et al. [2014],	Não	Não	Sim	Não	Não	Não
Padhi and Mohapatra [2011],	Não	Não	Não	Não	Não	Não
Souza and Pereira [2009],	Não	Não	Sim	Sim	Não	Não
Arief et al. [2016],	Não	Não	Não	Não	Não	Não
Domingos et al. [2016]	Não	Não	Não	Não	Não	Não
Ralha and Silva [2012]	Não	Não	Sim	Não	Não	Não
Erven et al. [2017]	Sim	Não	Sim	Não	Não	Não
Fraga et al. [2017]	Não	Não	Sim	Sim	Não	Não
Davydenko et al. [2017]	Sim	Não	Não	Não	Não	Não
Melo and Ferreira [2016]	Sim	Não	Sim	Não	Sim	Não
Carvalho et al. [2013]	Não	Não	Não	Não	Não	Não
O presente trabalho.	Sim	Sim	Sim	Não	Sim	Sim

4- Metodologia

Neste capítulo, o desenho metodológico desta dissertação é apresentado. A Figura 6 Ilustra o fluxograma da metodologia desenvolvida. Uma série de nove etapas estão agrupadas por três cenários distintos. No cenário 1, o objetivo é selecionar órgãos cuja frequência de participações das empresas em suas licitações sejam discrepantes. Para tanto, são previstas as seguintes etapas: na etapa 1, os lotes de licitações são classificados; na etapa 2, são gerados grafos bipartidos a partir dos dados dos lotes de licitações, empresas e órgãos públicos; na etapa 3, o valor de entropia dos vértices de órgãos é calculada; na etapa 4, os órgãos com o menor valor de entropia são selecionados. No cenário 2, o objetivo é identificar grupos de empresas com indícios de conluio. Para tanto, são previstas as seguintes etapas: na etapa 5 é um gerado um grafo com todas as empresas licitantes; na etapa 6, são detectadas comunidades de empresas com o método GN; na etapas 7, são detectadas comunidades de empresas com o método CNM; na etapas 8, são selecionadas as comunidades com maior coesão. O Cenário 3 reúne informações produzidas nos cenários 1 e 2 em um único grafo.

Figura 6 – Fluxograma da metodologia desenvolvida



As seções seguintes desenvolvem os cenários e etapas da metodologia desenvol-

vida da seguinte maneira: na Seção 4.1, os conceitos básicos de redes e a notação usada ao longo do trabalho são introduzidos, e as duas abordagens centradas em mineração em grafos são apresentadas. Na Seção 4.2, é apresentada a estratégia de validação dos experimentos.

4.1- Redes: conceitos básicos e notação

Ao longo do trabalho, denotamos $G = (V, E, w)$ como uma rede simples, não direcionada e ponderada com um conjunto de nós V de cardinalidade n e o conjunto de arestas E de cardinalidade m. Se há uma aresta (v_i, v_j) conectando vértices v_i e v_j, dizemos que estes vértices são adjacentes. Um peso $w_{ij} > 0$ é atribuído se os vértices v_i, v_j forem adjacentes. Abordamos o problema de identificação de conluio ou restrição à competitividade criando dois tipos grafos que serão construídos na parte de experimentos computacionais. Importante ressaltar que para cada CNAE, será construído um grafo do mesmo tipo descrito no Cenário 1. Assim, o total de grafos construídos com as características do Cenário 1 são exatamente iguais ao número de CNAEs estudados neste trabalho. As características dos grafos dos dois cenários são descritos nas seções a seguir. A Seção 4.1.1 aborda os grafos gerados para o Cenário 1, a Seção 4.1.2 aborda o grafo gerado para o Cenário 2 e a Seção 4.1.3 aborda o grafo gerado para o Cenário 3. Para todos os cenários, a geração dos grafos é feita com base nos dados estão estruturados de acordo com o modelo relacional apresentado na Seção 2.1.1, mais especificamente no modelo relacional apresentado na Figura 1.

4.1.1 Cenário 1: redes bipartidas de órgãos e empresas

Neste cenário, seja $G_1 = (V_1, E_1, w)$ um grafo bipartido, tal que $V_1 = V_e \cup V_o$ e $V_e \cap V_o = \emptyset$. Os nós do conjunto V_e representam as empresas privadas e cada nó em V_o representa um órgão público. Um vértice $v_i \in V_e$ está conectado a um vértice $v_j \in V_o$, se a empresa privada v_i participou de uma licitação realizada pelo órgão público v_j durante

o período de tempo analisado, e o peso w_{ij} atribuído a essa aresta é dado pelo número de participações de v_i em licitações realizadas pelo órgão v_j.

Tendo em vista que a frequência de participação de uma empresa em licitações pode ser influenciada pelo tipo de produto no lote que está sendo adquirido, os grafos G_1 serão agrupados pela classificação do lote. Para tanto, usaremos a classificação CNAE, discutida na Seção 2.1.3, para classificar os lotes. Os lotes das licitações serão classificados de acordo com a classe da CNAE das empresas que dela participam. Tendo em vistas as características da CNAE apresentadas na Seção 2.1.3, usaremos até o quarto nível mais específico da classificação para agrupar os lotes de licitação . A "classe" é o quarto nível mais específico de classificação da CNAE. Ela é representada por cinco dígitos numéricos do sistema de codificação, conforme indicado na Tabela 5. Isso significa que todas as empresas que participaram de um mesmo lote e possuem o mesmo número CNAE, considerado até o seu quarto nível, serão agrupadas para então formar uma rede de relacionamento entre os órgãos das licitações e as respectivas empresas.

O Algoritmo 2 apresenta como os lotes das licitações são classificados pelo CNAE. Observe que o lote é a entrada do Algoritmo 2 representado pela variável *lot*. A saída deste algoritmo é um vetor de CNAEs com as classificações para aquele lote. O lote terá tantas classificações de CNAE quantas CNAEs em comum tiverem as empresas. Caso as empresas não possuam nenhum CNAE em comum, o lote não terá classificação.

Tabela 5 – Organização hierárquica do CNAE

Nível	**Identificação**
1º	Código alfabético de 1 dígito
2º	Código numérico de 2 dígitos
3º	Código numérico de 3 dígitos
4º	**Código numérico de 4 dígitos + DV**
5º	Código numérico de 7 dígitos (incluindo o DV)

O Algoritmo 3 apresenta como cada grafo G_1 do Cenário 1 é construído. O Algoritmo 4 retorna a entropia de cada vértice do grafo. Observe que o número CNAE é uma entrada do Algoritmo 3 e para cada CNAE um grafo é construído. Observe que: $selectCompanies(ncnae)$ é a função que retorna todas as empresas privadas com um determinado CNAE dado como entrada; $selectPublicAgencies(ncnae)$ retorna o conjunto de todos os órgãos públicos com licitações associadas a um determinado CNAE durante o período de tempo analisado; $nbids(v_i, v_j, ncnae)$ é a função que retorna o número de lotes de licitações que a empresa v_i participou no órgão v_j associado ao CNAE de número

Algoritmo 2 – classifyLot(lot)

Output: $cnaesByLote[\,]$: a cnaes vector

1: $companies \leftarrow selectCompaniesByLot(lot)$
2: $ncnaes \leftarrow selectCnaesClassByCompany(nCompanies[0])$
3: $cnaesByLot \leftarrow ncnaes$
4: **for** $nc \in ncnaes$ **do**
5: **for** $cp \in companies$ **do**
6: $cnaesByLot \leftarrow cnaesByLot \cap selectCnaesClassByCompany(cp)$
7: **end for**
8: **end for**
9: **return** $cnaesByLot$

$ncnae$ para os lotes que foram classificados pelo Algoritmo 2.

Algoritmo 3 – generateGraph1($ncnae$)

1: $V_e \leftarrow selectCompanies(ncnae)$; $V_o \leftarrow selectPublicAgencies(ncnae)$
2: $E \leftarrow \emptyset$ $V \leftarrow V_e \cup V_o$
3: **for** $v_i \in V_e$ **do**
4: **for** $v_j \in V_o$ **do**
5: $edgeW \leftarrow nbids(v_i, v_j, ncnae)$
6: **if** $edgeW \geq 1$ **then**
7: $E \leftarrow E \cup \{v_i, v_j\}$
8: $w_{ij} \leftarrow edgeW$
9: **end if**
10: **end for**
11: **end for**
12: **return** G

A saída do Algoritmo 3 é fornecida como uma entrada para o Algoritmo 4, que

calcula a entropia de cada vértice do grafo bipartido.

Algoritmo 4 – computeEntropy(G_1)

1: $result \leftarrow$ [ ]
2: $V_o \leftarrow getNodes(G_1)$
3: **for** $v \in V_o$ **do**
4: $E_v \leftarrow$ set of edges in $E(G_1)$ incident to v
5: $SumW_v \leftarrow \sum_{(i,j)\in E_v} w_{ij}$
6: $s \leftarrow \sum_{(i,j)\in E_v}((w_{ij}/SumW_v)\ \log_2(SumW_v/w_{ij}))$
7: $logN \leftarrow 0$
8: **if** $\deg(v) > 1$ **then**
9: $logN \leftarrow 1/(\log_2 \deg(v))$
10: **end if**
11: $H_v \leftarrow s \times logN$
12: $result[v] \leftarrow H_v$
13: **end for**
14: **return** $result$

4.1.2 Cenário 2: rede de empresas

Nesta seção, a construção do grafo de empresas que participaram de todas as licitações de um dado período observado é apresentada. O grafo construído é valorado e o peso de cada aresta é definido. Esta rede define o grafo de empresas para o Cenário 2 e é construída pelo Algoritmo 5. Os algoritmos de identificação de comunidades do grafo e de ranqueamento das empresas são apresentados.

No Cenário 2, seja $G_2 = (V_2, E_2, w')$ um grafo em que o conjunto de nós V_2 representa todas as empresas privadas que participaram de todas as licitações ofertadas por todos os órgãos estudados no Cenário 1. Note que G_2 é uma rede de empresas privadas, de modo que duas delas estão conectadas se tiverem participado da mesma licitação disputando pelo mesmo lote. Um peso w'_{ij} é atribuído a cada aresta (v_i, v_j) conectando duas empresas privadas v_i e v_j de acordo com a seguinte expressão:

$$w'_{ij} = na_{i,j}fa + nb_{i,j}fb + nr_{i,j}fr + nc_{i,j}fc + nt_{i,j}ft + ne_{i,j}fe + nm_{i,j}fm + np_{i,j}fp, \quad (7)$$

onde nr_{ij} é o número de sócios com grau de parentesco entre as empresas v_i e v_j; $na_{ij}, nb_{ij}, nr_{ij}, nc_{ij}, nt_{ij}, ne_{ij}$ e nm_{ij} são o número de endereços, o número de lotes em licitações, número de sócios, número de contadores, número de telefones, número de funcionários e número de endereços de e-mail em comum entre as empresas v_i e v_j, respectivamente. Conforme já disposto na Seção 2.1.2, é razoável supor que, se duas empresas privadas têm os atributos descritos acima em comum e participam das mesmas licitações, elas podem estar agindo em conluio para simular uma concorrência na licitação pública. Para cada um destes atributos foi designado um peso. A expressão para os pesos foi obtida de acordo com a opinião dos especialistas da área. Os fatores são definidos da seguinte forma:

$$
\begin{aligned}
fa &= \frac{\sum_{(v_i,v_j)\in E_2} nb_{ij}}{\sum_{(v_i,v_j)\in E_2} na_{ij}}, \\
fb &= 1, \\
fr &= \frac{\sum_{(v_i,v_j)\in E_2} nb_{ij}}{\sum_{(v_i,v_j)\in E_2} nr_{ij}}, \\
fc &= \frac{\sum_{(v_i,v_j)\in E_2} nb_{ij}}{\sum_{((v_i,v_j))\in E_2} nc_{ij}}, \\
ft &= \frac{\sum_{(v_i,v_j)\in E_2} nb_{ij}}{\sum_{(v_i,v_j)\in E_2} nt_{ij}}, \\
fe &= \frac{\sum_{(v_i,v_j)\in E_2} nb_{ij}}{\sum_{(v_i,v_j)\in E_2} ne_{ij}}, \\
fm &= \frac{\sum_{(v_i,v_j)\in E_2} nb_{ij}}{\sum_{(v_i,v_j)\in E_2} nm_{ij}}, \\
fp &= \frac{\sum_{(v_i,v_j)\in E_2} nb_{ij}}{\sum_{(v_i,v_j)\in E_2} np_{ij}},
\end{aligned}
\tag{8}
$$

Os Algoritmos 5, 6 e 7 apresentam como o grafo G_2 do Cenário 2 é construído e as comunidades do grafo são detectadas e ranqueadas. Observe que o Algoritmo 6 usa o método GN e o Algoritmo 7 usa o método CNM para detecção de comunidades. A saída do Algoritmo 5 é fornecida como uma entrada dos Algoritmos 6 e 7.

A função $girvanNewmam(G_2)$ retorna um vetor de partições p para o grafo G_2, de forma que cada partição p_i é composta por k_i comunidades denotadas por $C_{k_i,i}$. A função $modularity(p_i, G_2)$ calcula a modularidade de uma dada partição p_i para o grafo G_2. A função $max(p)$ retorna a partição p_{max} de maior modularidade para o grafo G_2; A função $ratio\_w\_n\_e(G_2, C_{k_i,i})$ calcula a razão da soma dos pesos sob o número das

Algoritmo 5 – generateGraph2()

```
1: V ← selectCompanies()
2: for v_i ∈ V do
3:    for v_j ∈ V do
4:       nb ← nbids(v_i, v_j)
5:       if nb ≥ 1 then
6:          ne ← employeesBetween(v_i, v_j)
7:          nt ← phonesBetween(v_i, v_j)
8:          np ← partnersBetween(v_i, v_j)
9:          nm ← emailsBetween(v_i, v_j)
10:         nc ← accountantBetween(v_i, v_j)
11:         na ← addressBetween(v_i, v_j)
12:         nr ← relativesBetween(v_i, v_j)
13:         E ← E ∪ {v_i, v_j}
14:         w_ij ← weigth(ne, nt, np, nm, nc, na, nr)
15:      end if
16:   end for
17: end for
18: return G
```

arestas que conectam os vértices pertencentes às comunidades $C_{k_i,i}$ de p_{max}. A função $sort\_descending(result)$ retorna em ordem não-crescente os valores do vetor $result$.

A função $clausetNewmanMoore(G_2)$ retorna partição p_{cnm} de maior modularidade para o grafo G_2, de acordo com o método CNM; a função $ratio\_w\_n\_e(G_2, C_{k_i,i})$ calcula a razão da soma dos pesos sob o número das arestas que conectam os vértices pertencentes às comunidades $C_{k_i,i}$ de p_{max}; a função $sort\_descending(result)$ retorna em ordem não-crescente os valores do vetor $result$.

Algoritmo 6 – detectCollusionCompaniesGN(G_2)

```
Input: G_2 = G2(V_2, E_2, w')
Output: result ← [ ]
1: p ← girvanNewmam(G_2)
2: q ←[ ]
3: for p_i ∈ p do
4:    q[p_i] ← modularity(p_i, G_2)
5: end for
6: p_max ← max(q)
7: for C_{k_i,i} ∈ p_max do
8:    result[C_{k_i,i}] ← ratio_w_n_e(G_2, C_{k_i,i})
9: end for
10: result ← sort_descending(result)
11: return result
```

Algoritmo 7 – detectCollusionCompaniesCNM(G_2)

Input: $G_2 = G2(V_2, E_2, w\prime)$
Output: $result \leftarrow [\,]$
1: $p_{cnm} \leftarrow clausetNewmanMoore(G_2)$
2: **for** $C_i \in p\_cnm$ **do**
3: $\quad result[C_i] \leftarrow ratio\_w\_n\_e(G_2, C_i)$
4: **end for**
5: $result \leftarrow sort\_descending(result)$
6: **return** $result$

4.1.3 Cenário 3: grafo de consolidação dos Cenários 1 e 2

Os órgãos de controle tem atuação restrita sobre os órgãos da administração pública. Portanto, uma relação de empresas em conluio só lhes é útil se também forem identificados os órgãos da administração pública em que tais empresas atuam. Para relacionar os órgãos da administração pública e as comunidades de empresas com indícios de atuação em conluio, propomos uma modelagem via grafos reunindo os resultados do Cenário 1 e do Cenário 2 em um único grafo de maneira a permitir que o auditor, a partir da sua experiência, expertise e outros critérios, possa selecionar o órgão que deve ser auditado.

No Cenário 3, seja $G_3 = (V_3, E_3, w'')$ um grafo bipartido, tal que $V_3 = V_{e'} \cup V_{o'}$ e $V_{e'} \cap V_{o'} = \emptyset$. Os nós do conjunto $V_{e'}$ representam as empresas privadas das dez comunidades com maior coesão identificadas no Cenário 3. Os vértices do conjunto $V_{o'}$ representam órgãos públicos cujas licitações as empresas em $V_{e'}$ tenham participado. Um vértice $v_{i'} \in V_{e'}$ está conectado a um vértice $v_{j'} \in V_{o'}$, se a empresa privada $v_{i'}$ participou de uma licitação realizada pelo órgão público $v_{j'}$ durante o período de tempo analisado, e o peso $w''_{i'}$ atribuído a essa aresta é dado pelo número de participações de $v_{i'}$ em licitações realizadas pelo órgão $v_{j'}$.

4.2- Validação

Mensurar os resultados de predição dos classificadores é possível em abordagens onde a amostra de dados está rotulada. Uma abordagem dessas não é possível no presente trabalho, pois a amostra de dados não está rotulada. Não é possível selecionar uma amostra do universo das contratações e realizar auditorias de fiscalização a fim de rotulá-las, uma vez que isso é economicamente inviável. Conforme explicado na Seção 2.1.2, não há informações a respeito de qual a proporção de casos com irregularidades do universo total das contratações, não há rótulos adequados identificando quais as contratações são irregulares e que tipo de irregularidade ocorre. Também não é possível iniciar um trabalho de rotulagem dos dados, já que seria necessário realizar várias auditorias, e isso tem um custo proibitivo de realização. Por esses motivos, recorre-se a duas abordagens para avaliação dos resultados encontrados: avaliação empírica de casos descobertos e a criação de redes com dados sintéticos. A avaliação empírica será feita a partir de um detalhamento dos casos encontrados. A criação de dados sintéticos depende da identificação de padrões de formação das redes produzidas pelo Cenário 1. A ideia é reproduzir o comportamento dessas redes a partir da identificação de que essas redes em geral seguem o modelo da lei de potências, por exemplo. Para identificar esses padrões realizamos testes estatísticos para a distribuição de graus dos vértices dos grafos do cenário 1 de acordo como descrito na seção a seguir.

4.2.1 Verificação de padrões nos Grafos do cenário 1

O objetivo desta etapa é definir um valor de corte (H_v) para o valor de entropia. O valor de corte separa o conjunto de órgãos que deve passar por uma análise mais detida de suas licitações e os órgãos em que não serão realizados nenhuma ação de fiscalização adicional. Para tanto, verificaremos se os grafos G_1 seguem algum padrão a fim de gerarmos dados sintéticos a partir dos padrões identificados.

Usaremos o teste de K-S, que foi apresentado na Seção 2.4, para verificar se os dados estatísticos dos grafos G_1 são aderentes a distribuições estatísticas. Para cada

grafo G_1, testaremos a aderência de distribuições estatísticas aos seguintes conjunto de dados:

- da sequência de graus dos vértices do conjunto V_1;
- da sequência de graus dos vértices do conjunto V_o;
- da sequência de graus dos vértices do conjunto V_e;
- da sequência de pesos das arestas do conjunto E_1.

Esses conjuntos de dados serão testadas para todas as seguintes distribuições: “alpha”, “anglit”, “arcsine”, “beta”, “betaprime”, “bradford”, “burr”, “cauchy”, “chi”, “chi2”, “cosine”, “dgamma”, “dweibull”, “erlang”, “expon”, “exponnorm”, “exponweib”, “exponpow”, “f”, “fatiguelife”, “fisk”, “foldcauchy”, “foldnorm”, “frechet r”, “frechet l”, “genlogistic”, “genpareto”, “gennorm”, “genexpon”, “genextreme”, “gausshyper”, “gamma”, “gengamma”, “genhalflogistic”, “gilbrat”, “gompertz”, “gumbel r”, “gumbel l”, “halfcauchy”, “halflogistic”, “halfnorm”, “halfgennorm”, “hypsecant”, “invgamma”, “invgauss”, “invweibull”, “johnsonsb”, “johnsonsu”, “ksone”, “kstwobign”, “laplace”, “levy”, “levy l”, “levy stable”, “logistic”, “loggamma”, “loglaplace”, “lognorm”, “lomax”, “maxwell”, “mielke”, “nakagami”, “ncx2”, “ncf”, “nct”, “norm”, “pareto”, “pearson3”, “powerlaw”, “powerlognorm”, “powernorm”, “rdist”, “reciprocal”, “rayleigh”, “rice”, “recipinvgauss”, “semicircular”, “t”, “triang”, “truncexpon”, “truncnorm”, “tukeylambda”, “uniform”, “vonmises”, “vonmises line”, “wald”, “weibull min”, “weibull max” e “wrapcauchy”.

Portanto, testaremos para cada grafo G_1 se o conjunto de dados é aderente a alguma das 89 (oitenta e nove) distribuições estatísticas relacionadas. A verificação de aderência com o teste de K-S será realizada com para as seguintes hipóteses nula e alternativa abaixo de acordo com o conjunto em análise:

(A) Para a sequência de graus dos vértices do conjunto V_1:

- H_0: A sequência de graus dos vértices da rede segue a distribuição testada.
- H_1: A sequência de graus dos vértices da rede não segue a distribuição testada.
- O nível de significância é $\alpha = 0.05$.

(B) Para a sequência de graus dos vértices do conjunto V_o:

- H_0: A sequência de graus dos vértices da partição de V_o da rede segue a distribuição testada.
- H_1: A sequência de graus dos vértices da partição de V_o da rede não segue a distribuição testada.
- O nível de significância é $\alpha = 0.05$.

(C) Para a sequência de graus dos vértices do conjunto V_e:

- H_0: A sequência de graus dos vértices da partição de V_e da rede segue a distribuição testada.
- H_1: A sequência de graus dos vértices da partição de V_e da rede não segue a distribuição testada.
- O nível de significância é $\alpha = 0.05$.

(D) Para a sequência de pesos das arestas do conjunto E_1:

- H_0: A sequência de pesos das arestas do conjunto E_1 da rede segue a distribuição testada.
- H_1: A sequência de pesos das arestas do conjunto E_1 da rede não segue a distribuição testada.
- O nível de significância é $\alpha = 0.05$.

5- Experimentos Computacionais

Com a metodologia apresentada no Capítulo 4 foram realizados experimentos sobre o conjunto de dados de licitações de bens e serviços de órgãos do Estado do Rio de Janeiro. Os resultados dos experimentos são detalhados nas seções a seguir, onde na Seção 5.1, os recursos computacionais utilizados são apresentados; na Seção 5.2 o conjunto de dados é apresentado; a Seção 5.3 apresenta os resultados dos experimentos computacionais do Cenário 1, a Seção 5.4 apresenta os experimentos computacionais do Cenário 2, a Seção 5.5 apresenta um novo grafo com os resultados das duas abordagens.

5.1- Recursos utilizados

Implementamos todos os algoritmos desta pesquisa em Python. Usamos o pacote networkx e seus métodos para criar e manipular os grafos trabalhados. Foi utilizado o software Gephi para produzir as ilustrações dos grafos. Para computar os experimentos, foi utilizado um computador equipado com processador Intel Core I7 4790 3.60ghz de 4ª geração e 8 GB de memória.

5.2- Descrição do conjunto de dados

O conjunto de dados analisado corresponde às licitações de bens e serviços de 84 (oitenta e quatro) órgãos do Estado do Rio de Janeiro realizadas entre 2010 e 2018. Esses órgãos realizaram no período analisado 8.519 (oito mil quinhentos e dezenove) licitações, para a aquisição de 60.047 (sessenta mil e quarenta e sete) lotes. Participaram destas licitações 6.390 (seis trezentos e noventa) empresas.

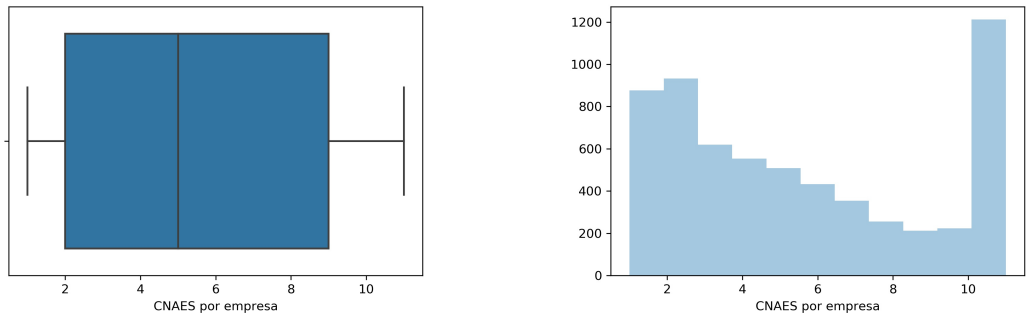
A Tabela 6 e os gráficos das Figuras 7, 8, 9 e 10 apresentam dados estatísticos do número de CNAEs por empresa, do número de licitações realizadas por órgão, do

número de lotes por licitação e do número de participantes por licitação. Essas estatísticas influenciam o tamanho e a topologia dos grafos G_1 e G_2.

Tabela 6 – Estatísticas de licitações por órgão, lotes por licitação, participantes por licitação de CNAEs por empresa do conjunto de dados analisado

	Mínimo	Máximo	Média	Mediana	Cerca superior
licitações por órgão	1	991	101,4	50,5	293
lotes por licitação	1	322	7	2	11
participantes por licitação	3	40	6,2	5	14
CNAEs por empresa	1	11	5,5	5	-

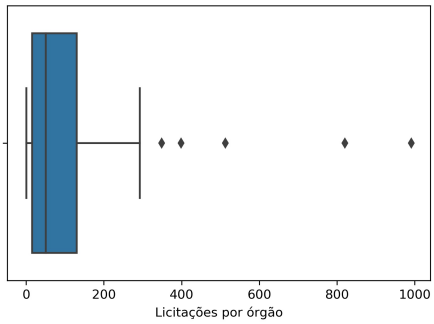
Pela análise do boxplot da Figura 7 (a) e do histograma da Figura 7 (b), verifica-se que, apesar de 50% (cinquenta por cento) das empresas terem entre 2 (dois) e 9 (nove) CNAEs, um número considerável de empresas, cerca de 1.200 (mil e duzentos), ou quase um quinto do número total de empresas, tem 10 (dez) CNAEs. Esse é um padrão esperado, tendo em vista a discussão apresentada na Seção 2.1.2. Algumas empresas se especializam em fornecer bens e produtos para administração pública e buscam ter diversos CNAEs para participarem de um número maior de licitações.



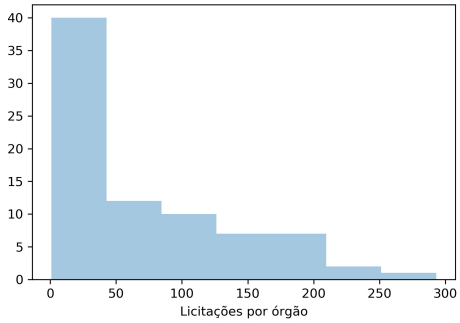
(a) (b)

Figura 7 – Número de CNAEs por empresa

Na Figura 8, são apresentados um boxplot e um histograma relativos ao número de licitações organizadas por órgão. No boxplot da Figura 8 (a) verifica-se que alguns poucos órgãos superam em grande quantidade o número de licitações realizadas pelos demais órgãos. Cerca de 40 (quarenta), dos 84 (oitenta e quatro) órgãos realizaram de 1 (uma) a 50 (cinquenta) licitações, enquanto 5 (cinco) órgãos realizaram mais de 300 (trezentas) licitações cada um.



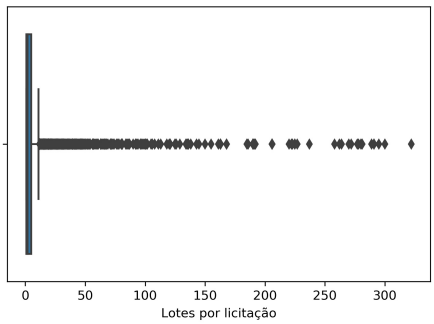
(a) Número de licitações realizadas por órgão



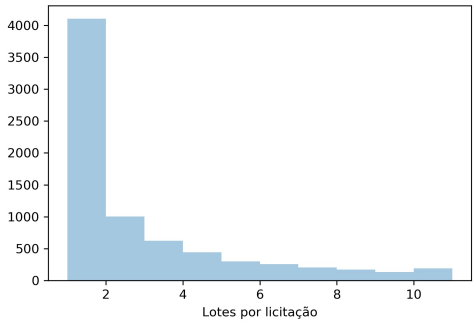
(b) Número de licitações realizadas por órgão sem os outliers

Figura 8 – Número de licitações realizadas por órgão

Na Figura 9, são apresentados um boxplot e um histograma relativos ao número de lotes por licitação. Pela análise do histograma da Figura 9 (b), quase metade das licitações, cerca de 4.000 (quatro mil) das 8.519 (oito mil quinhentos e dezenove), tem somente um lote por licitação.



(a) Número de lotes por licitação

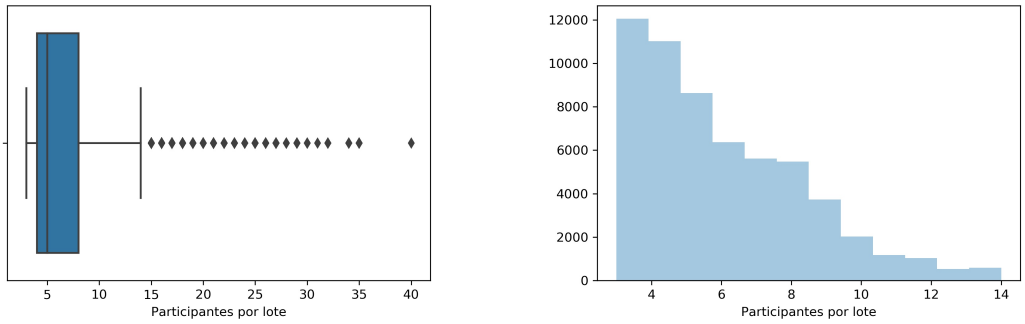


(b) Número de lotes por licitação sem os outliers

Figura 9 – Número de lotes por licitação

Na Figura 10 são apresentados um boxplot e um histograma relativos ao número de empresas participantes por lotes de licitação. O número de empresas participantes por lote de licitação é o que tem maior impacto no tamanho dos grafos gerados para os Cenários 1 e 2. Verifica-se que as licitações tem de 3 (três) a 14 (quatorze) empresas licitantes. Algumas poucas licitações tem mais do que 14 (quatorze) empresas licitantes. Um certamente só é válido se tiver ao menos 3 (três) licitantes. Pela análise do histograma da Figura 10 (10b), verifica-se que, dos 60.047 (sessenta mil e quarenta e sete) lotes de licitação, cerca de 12.000 (doze mil) lotes de licitação só tiveram o número mínimo de

empresas. Conforme discutido na Seção 2.1.1, esse padrão é o oposto do desejado. Seria desejável que a quantidade de empresas participantes por certame fosse concentrada em números superiores ao mínimo requerido. Conforme discutido na Seção 2.1.1, quanto maior o número de empresas licitantes, maior é a competitividade do certame.



(a) Número de empresas participantes por lote de licitação

(b) Número de empresas participantes por lote de licitação sem os outliers

Figura 10 – Número de empresas privadas participantes por lote de licitação

Na análise de fundo sobre a base de decisões do Tribunal de Contas do Estado do Rio de Janeiro (TCE-RJ), identificamos que o órgão fiscalizador sancionou 4 (quatro) dos 84 (oitenta) órgãos e 4 (quatro) das 6.390 empresas licitantes por uma série de irregularidades em licitações e contratações realizadas. Entre as irregularidades encontradas se destacam o conluio e conivência dos gestores para perpetuação de grupos empresariais com a restrição de competitividade nas contratações e licitações realizadas pelos órgãos licitantes.Esses casos foram rotulados na base de dados a fim de permitir um análise mais detida desses casos. Convém destacar que para os demais órgãos não há rótulo indicando se há ou não irregularidades em suas licitações. Conforme já esclarecido na Seção 1, essa base de dados não possui rótulos para classificar a maior parte dos órgãos quanto a irregularidades em suas licitações. Todos os casos aqui tratados foram anonimizados.

5.3- Resultados dos experimentos computacionais para o Cenário 1

Aplicamos o método desenvolvido na Seção 4.1 aos dados descritos na Seção 5.2. A partir do Algoritmo 2 de classificação dos lotes pelo CNAE, os lotes das licitações

foram classificados em 122 (cento e vinte e duas) CNAEs distintas. Usando aquela abordagem foi possível classificar 12.421 lotes de licitação. Cerca de 20,68% dos lotes de licitações foram classificados. Dos 84 (oitenta e quatro) órgãos presentes na base de dados, foi possível classificar lotes de 76 órgãos distintos. Destes lotes, cerca de 76% (setenta e seis por cento) foram classificados numa das seguintes CNAEs: 46443 “Comércio atacadista de produtos farmacêuticos para uso humano e veterinário”, 46451 “Comércio atacadista de instrumentos e materiais para uso médico, cirúrgico, ortopédico e odontológico” e 46397 “Comércio atacadista de produtos alimentícios em geral”, conforme ilustra a Figura 11.

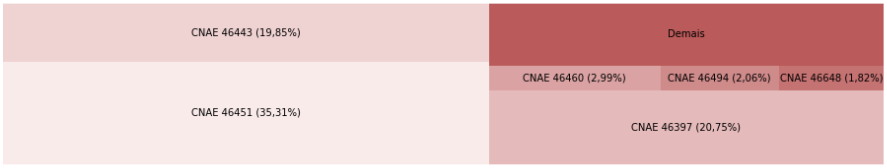


Figura 11 – Resultado da classificação dos lotes pelo Algoritmo 2

Utilizamos o Algoritmo 3 para gerar os grafos bipartido do Cenário 1. Cada grafo G_1 representa um CNAE. Foram gerados 122 (cento e vinte e dois) grafos. Os gráficos da Figura 12 reúnem informações estatísticas a respeito dos grafos gerados. São apresentadas o número de vértices, de arestas, dos vértices do conjunto V_o, que representam órgãos públicos, e dos vértices do conjunto V_e, que representam as empresas privadas.

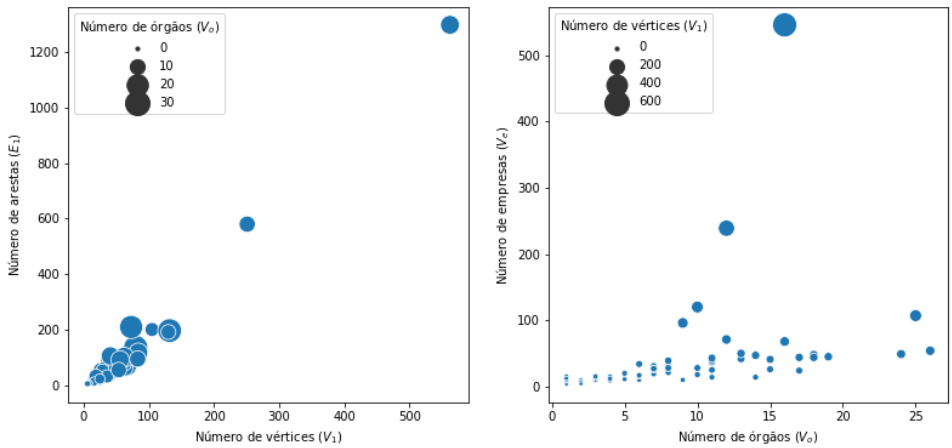


Figura 12 – Estatística descritiva dos grafos Cenário 1

De todos os grafos G_1 gerados pelo Cenário 1, destacamos o maior grafo em número de arestas e vértices nas Figuras 14 e 13. Ambas Figuras 14 e 13 reproduzem o grafo G_1^{46451}, que mostra as relações órgãos-empresas no conjunto de dados dos lotes classificados com o CNAE de número $ncnae = 46451$ (“comercio atacadista de instrumentos e materiais para uso medico cirúrgico ortopédico e odontológico”). A Figura 13 apresenta um layout que destaca as duas partições do grafo. Do lado esquerdo estão os vértices do conjunto V_e (empresas) e do lado direito os vértices do conjunto V_o (órgãos públicos). A Figura 14 apresenta um layout que destaca as conexões dos vértices do conjunto V_o com os vértices do conjunto V_e. Nas duas Figuras, os nós vermelhos são os vértices do conjunto V_o e os nós azuis são os vértices do conjunto V_e . O tamanho dos vértices do conjunto V_o é inversamente proporcional ao valor de entropia, denotado pela equação $\frac{1}{H_v}$. O grafo G_1^{46451} é formado 562 vértices e 1.297 arestas. O conjunto V_o possui 16 vértices. O conjunto V_e possui 546 vértices. A densidade desse grafo é dada por $d_{G_1^{46451}} \approx 0,0082$.

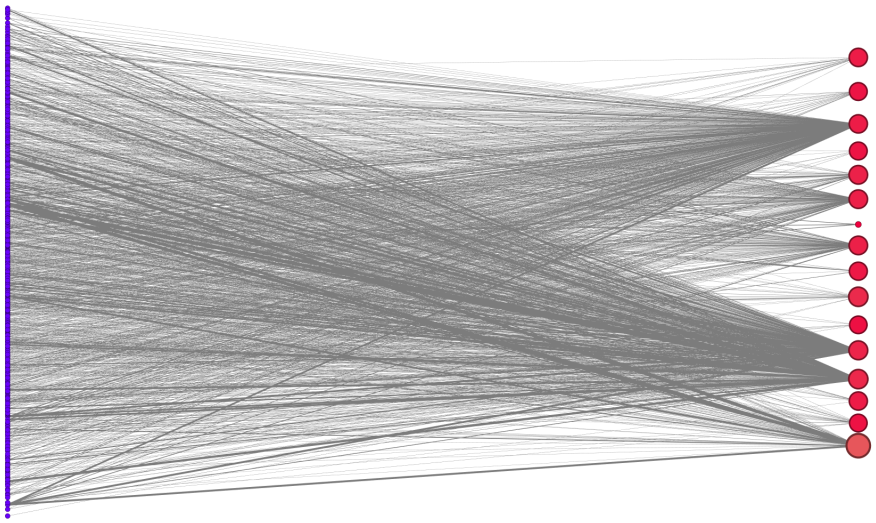


Figura 13 – Rede G_1^{46451} (“comercio atacadista de instrumentos e materiais para uso medico cirúrgico ortopédico e odontológico”)

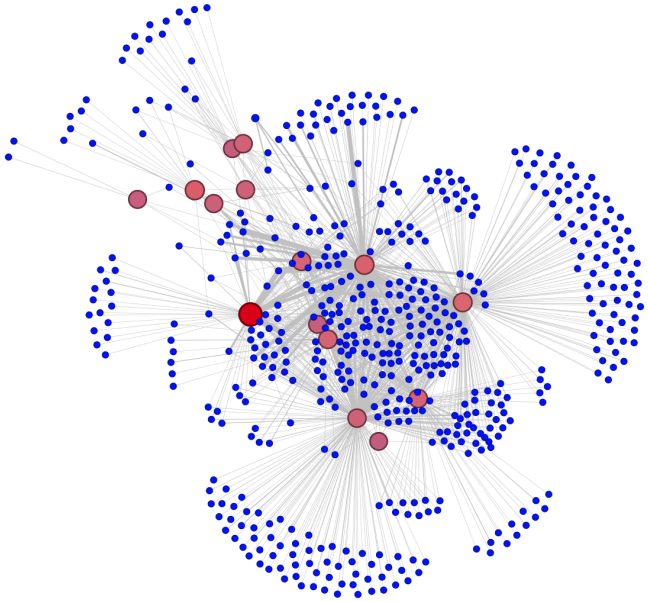


Figura 14 – Rede G_1^{46451} (“comercio atacadista de instrumentos e materiais para uso medico cirúrgico ortopédico e odontológico”)

Os gráficos de boxplot e histograma da Figura 15 apresentam os valores de entropia dos vértices do conjunto V_o do Grafo G_1^{46451}.

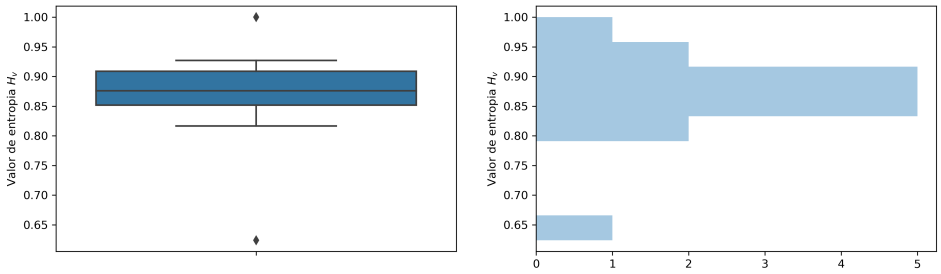


Figura 15 – Valor de entropia dos vértices do conjunto V_o do grafo G_1^{46451}

A Figura 16 reproduz o grafo G_1^{46451}, mas destacando o vértice do conjunto V_o com menor valor de entropia em H_{v_j} e seus vértices adjacentes.

O boxplot da Figura 17 representa o valor de entropia de todos os vértices dos dez grafos G_1 com maior número de órgãos. Nesse gráfico, destacamos os quatro órgãos para os quais havia rótulo de irregularidade. Pode se observar que os órgãos previamente rotulados com restrição a competitividade não formam nenhum padrão distinguível no gráfico. O valor de entropia dos órgãos previamente rotulados ficaram sempre acima do valor de entropia mediano em cada rede. Tendo em vista que os órgãos de controle somente selecionam uma pequena parcela da quantidade de órgãos para a realização

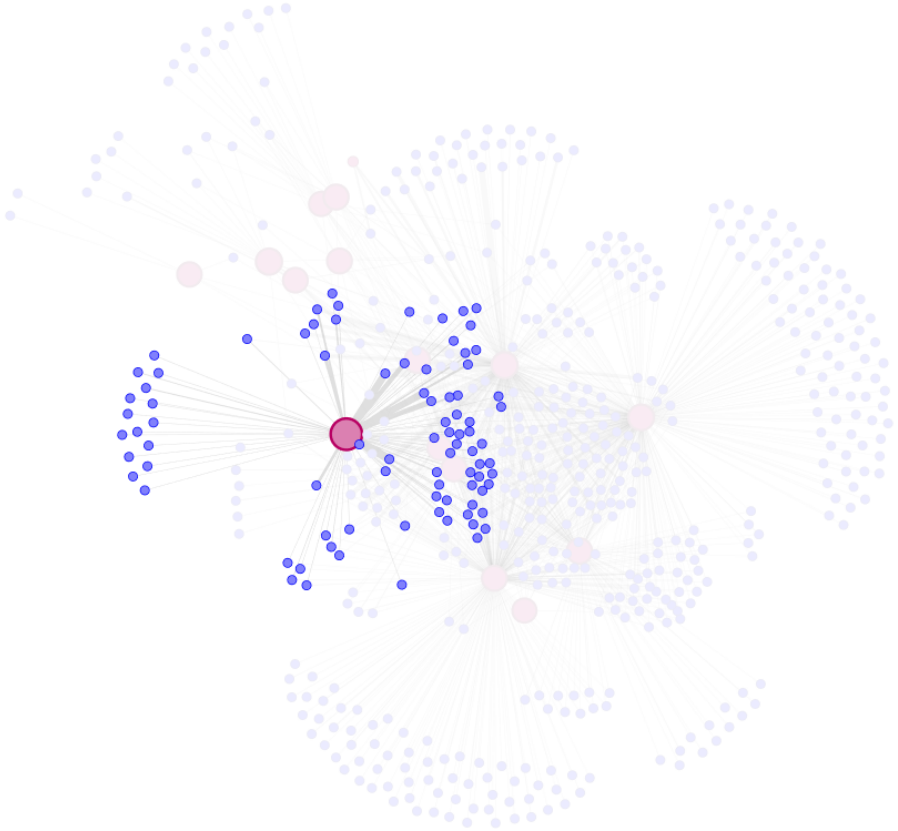


Figura 16 – Rede G_1^{46451} (“comercio atacadista de instrumentos e materiais para uso medico cirúrgico ortopédico e odontológico”, com o vértice de menor valor de entropia e seus adjacentes em destaque)

de auditorias, os órgão rotulados não seriam selecionados, já que tiveram valores de entropia sempre acima do valor mediano em cada rede.

5.3.1 Verificação de padrões nos Grafos do cenário 1

Seguindo a metodologia apresentada na Seção 4.2.1, usamos o teste de K-S apresentado na Seção 2.4 para testar se os conjuntos de dados de cada grafo G_1 eram aderentes a alguma daquelas distribuições estatísticas listadas. Aplicamos o teste de K-S às sequências de graus dos vértices do conjunto V_1, V_o e V_e e da sequência de pesos das arestas do conjunto E_1 de cada um dos 122 (cento e vinte e dois) grafos G_1 a fim de verificar ajustamento àquelas 89 (oitenta e nove) distribuições estatísticas listadas na Seção 4.2.1.

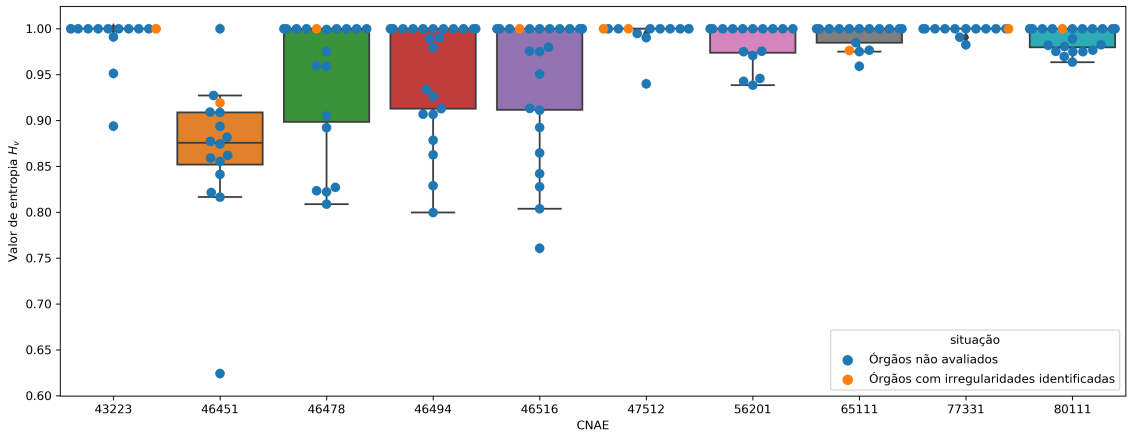


Figura 17 – Valor de entropia dos vértices do conjunto V_o dos dez grafos G_1 com maior número de órgãos

Foram executados, portanto, 43.432 (quarenta e três mil e quatrocentos e trinta e dois), testes K-S. As tabelas 7, 8, 9 e 10 apresentam os resultados dos testes para a sequência de graus dos vértices do conjunto V_e, do conjunto V_o, do conjunto V_1 e para a sequência de pesos das arestas dos grafos do Cenário 1, respectivamente. É apresentada somente a distribuição que teve o melhor ajuste ao conjunto de dados para cada grafo.

A Tabela 7 apresenta os resultados dos testes K-S para sequência de graus dos vértices do conjunto V_e dos grafos do Cenário 1 com mais de 6 (seis) órgãos. São apresentados somente o resultado da distribuição que teve o melhor ajuste ao conjunto de dados testado. A coluna “Rede $(ncnae)$” apresenta a CNAE do grafo correspondente. A coluna “N” apresenta o número de graus, número N, o tamanho da amostra ou o número de vértices do conjunto V_e. A coluna $D_{\alpha=0.05}(N)$ apresenta o valor $D_{\alpha}(N)$ ou kolmogorov crítico com nível de significância de 5% para o número de amostras correspondente. Esses valores foram recuperados da Tabela 3. A coluna “Nome da distribuição” apresenta o nome da distribuição estatística que teve o melhor ajuste, ou o menor valor de D, para a sequência de graus dos vértices do conjunto V_e testado. A coluna “D” apresenta os valores D para a amostra de dados testada. A coluna “Resultado do teste” apresenta o resultado final do teste. Caso o valor de D seja menor que o valor de $D_{\alpha=0.05}(N)$, a hipótese nula não é rejeitada. Ou seja, não há argumentos para rejeitar a hipótese de que a sequência de graus dos vértices do conjunto V_e segue a distribuição testada. Caso o valor de D seja maior que o valor de $D_{\alpha=0.05}(N)$, a hipótese nula é rejeitada. A hipótese nula foi rejeitada em todos os testes K-S aplicados para verificar se a sequência de graus

dos vértices do conjunto V_1 se ajustavam a alguma das distribuições listadas, a exceção das redes $G_1^{43291}, G_1^{65120}, G_1^{81222}$ e G_1^{82997} para as distribuições “ncf”, “levy”, “ncx2” e “ncf” respectivamente.

Tabela 7 – Resultado dos testes Kolmogorov-Smirnov da sequência de graus do conjunto de vértices V_e dos grafos do Cenário 1 com número de órgãos superior a 6.

Rede ($ncnae$)	N	$D_{\alpha=0.05}(N)$	Nome da distribuição	D	Resultado do teste
18113	28	0,250	invgamma	0,289	H_0 rejeitada
18130	42	0,205	levy	0,317	H_0 rejeitada
33121	31	0,238	rayleigh	0,413	H_0 rejeitada
43223	44	0,201	invweibull	0,368	H_0 rejeitada
43291	14	0,349	ncf	0,213	H_0 não rejeitada
45307	37	0,218	invweibull	0,368	H_0 rejeitada
46397	40	0,210	levy	0,317	H_0 rejeitada
46427	18	0,309	invgamma	0,331	H_0 rejeitada
46443	239	0,088	levy	0,317	H_0 rejeitada
46451	546	0,058	levy	0,317	H_0 rejeitada
46460	96	0,139	maxwell	0,301	H_0 rejeitada
46478	45	0,198	invweibull	0,368	H_0 rejeitada
46494	107	0,131	invgamma	0,314	H_0 rejeitada
46516	54	0,185	invgauss	0,335	H_0 rejeitada
46648	120	0,124	halfgennorm	0,340	H_0 rejeitada
47440	19	0,301	rayleigh	0,449	H_0 rejeitada
47512	48	0,192	rice	0,393	H_0 rejeitada
47610	25	0,264	rice	0,393	H_0 rejeitada
47890	50	0,188	genpareto	0,452	H_0 rejeitada
49230	27	0,254	rice	0,393	H_0 rejeitada
49302	47	0,194	invgauss	0,385	H_0 rejeitada
56201	68	0,165	invgamma	0,329	H_0 rejeitada
62015	31	0,238	genhalflogistic	0,462	H_0 rejeitada
62040	26	0,259	rayleigh	0,453	H_0 rejeitada
62091	27	0,254	invweibull	0,368	H_0 rejeitada
65111	24	0,269	genexpon	0,282	H_0 rejeitada
65120	14	0,349	levy	0,317	H_0 não rejeitada
69206	25	0,264	levy	0,317	H_0 rejeitada
77110	39	0,213	rice	0,393	H_0 rejeitada
77331	44	0,201	levy	0,317	H_0 rejeitada
77390	71	0,161	rice	0,393	H_0 rejeitada
80111	49	0,190	f	0,303	H_0 rejeitada
80200	21	0,287	levy	0,317	H_0 rejeitada
81222	26	0,259	ncx2	0,257	H_0 não rejeitada
81290	41	0,208	maxwell	0,300	H_0 rejeitada
82300	28	0,250	halflogistic	0,462	H_0 rejeitada
82997	10	0,409	ncf	0,235	H_0 não rejeitada
95118	43	0,203	rice	0,393	H_0 rejeitada

A Tabela 8 apresenta os resultados dos testes K-S para sequência de graus dos vértices do conjunto V_o dos grafos do Cenário 1 com mais de 6 (seis) órgãos. São apresentados somente o resultado da distribuição que teve o melhor ajuste ao conjunto de dados testado. Para esse teste N é o número de vértices do conjunto V_o.

Tabela 8 – Resultado dos testes Kolmogorov-Smirnov da sequência de graus do conjunto de vértices V_o dos grafos do Cenário 1 com número de órgãos superior a 6.

Rede ($ncnae$)	N	$D_{\alpha=0.05}(N)$	Nome da distribuição	D	Resultado do teste
18113	10	0,409	gennorm	0,449	H_0 rejeitada
18130	13	0,361	genlogistic	0,434	H_0 rejeitada
33121	7	0,483	nct	0,540	H_0 rejeitada
43223	17	0,318	lomax	0,448	H_0 rejeitada
43291	11	0,391	dgamma	0,522	H_0 rejeitada
45307	11	0,391	lomax	0,576	H_0 rejeitada
46397	11	0,391	f	0,562	H_0 rejeitada
46427	10	0,409	fatiguelife	0,589	H_0 rejeitada
46443	12	0,375	invgamma	0,567	H_0 rejeitada
46451	16	0,327	betaprime	0,367	H_0 rejeitada
46460	9	0,430	powerlognorm	0,547	H_0 rejeitada
46478	19	0,301	burr	0,334	H_0 rejeitada
46494	25	0,264	pareto	0,422	H_0 rejeitada
46516	26	0,259	betaprime	0,352	H_0 rejeitada
46648	10	0,409	powerlognorm	0,490	H_0 rejeitada
47440	7	0,483	f	0,614	H_0 rejeitada
47512	18	0,309	gengamma	0,474	H_0 rejeitada
47610	11	0,391	burr	0,490	H_0 rejeitada
47890	13	0,361	johnsonsu	0,457	H_0 rejeitada
49230	7	0,483	genexpon	0,561	H_0 rejeitada
49302	14	0,349	f	0,411	H_0 rejeitada
56201	16	0,327	genlogistic	0,413	H_0 rejeitada
62015	8	0,454	johnsonsu	0,540	H_0 rejeitada
62040	7	0,483	powerlognorm	0,510	H_0 rejeitada
62091	7	0,483	lomax	0,523	H_0 rejeitada
65111	17	0,318	invgamma	0,438	H_0 rejeitada
65120	14	0,349	invgamma	0,493	H_0 rejeitada
69206	8	0,454	chi	0,651	H_0 rejeitada
77110	8	0,454	mielke	0,632	H_0 rejeitada
77331	18	0,309	fatiguelife	0,322	H_0 rejeitada
77390	12	0,375	nct	0,480	H_0 rejeitada
80111	24	0,269	exponnorm	0,349	H_0 rejeitada
80200	8	0,454	genexpon	0,493	H_0 rejeitada
81222	15	0,338	burr	0,340	H_0 rejeitada
81290	15	0,338	betaprime	0,388	H_0 rejeitada
82300	8	0,454	fatiguelife	0,533	H_0 rejeitada
82997	9	0,430	f	0,433	H_0 rejeitada
95118	11	0,391	ncf	0,473	H_0 rejeitada

A Tabela 9 apresenta os resultados dos testes K-S para sequência de graus dos vértices dos grafos do Cenário 1 com mais de 6 (seis) órgãos. São apresentados somente o resultado da distribuição que teve o melhor ajuste ao conjunto de dados testado. Para esse teste N é o número de vértices dos grafos do cenário 1.

Tabela 9 – Resultado dos testes Kolmogorov-Smirnov da sequência de graus do conjunto de vértices dos grafos do Cenário 1 com número de órgãos superior a 6.

Rede ($ncnae$)	N	$D_{\alpha=0.05}(N)$	Nome da distribuição	D	Resultado do teste
18113	38	0,215	f	0,308	H_0 rejeitada
18130	55	0,183	levy	0,317	H_0 rejeitada
33121	38	0,215	invgauss	0,341	H_0 rejeitada
43223	61	0,174	invgamma	0,287	H_0 rejeitada
43291	25	0,264	levy	0,343	H_0 rejeitada
45307	48	0,192	invgauss	0,320	H_0 rejeitada
46397	51	0,190	f	0,305	H_0 rejeitada
46427	28	0,250	levy	0,317	H_0 rejeitada
46443	251	0,086	ncx2	0,311	H_0 rejeitada
46451	562	0,057	levy	0,317	H_0 rejeitada
46460	105	0,133	levy	0,317	H_0 rejeitada
46478	64	0,170	invgamma	0,262	H_0 rejeitada
46494	132	0,118	levy	0,317	H_0 rejeitada
46516	80	0,152	levy	0,317	H_0 rejeitada
46648	130	0,119	invgauss	0,321	H_0 rejeitada
47440	26	0,259	ncf	0,322	H_0 rejeitada
47512	66	0,167	levy	0,317	H_0 rejeitada
47610	36	0,221	alpha	0,317	H_0 rejeitada
47890	63	0,171	f	0,354	H_0 rejeitada
49230	34	0,227	invgamma	0,295	H_0 rejeitada
49302	61	0,174	invgamma	0,309	H_0 rejeitada
56201	84	0,148	f	0,311	H_0 rejeitada
62015	39	0,213	invweibull	0,377	H_0 rejeitada
62040	33	0,231	genexpon	0,343	H_0 rejeitada
62091	34	0,227	levy	0,317	H_0 rejeitada
65111	41	0,208	levy	0,317	H_0 rejeitada
65120	28	0,250	recipinvgauss	0,261	H_0 rejeitada
69206	33	0,231	powerlognorm	0,300	H_0 rejeitada
77110	47	0,194	levy	0,317	H_0 rejeitada
77331	62	0,172	levy	0,317	H_0 rejeitada
77390	83	0,149	invgauss	0,340	H_0 rejeitada
80111	73	0,159	powerlognorm	0,277	H_0 rejeitada
80200	29	0,246	pareto	0,276	H_0 rejeitada
81222	41	0,208	ncx2	0,210	H_0 rejeitada
81290	56	0,181	ncx2	0,242	H_0 rejeitada
82300	36	0,221	f	0,368	H_0 rejeitada
82997	19	0,301	foldcauchy	0,303	H_0 rejeitada
95118	54	0,185	invgauss	0,344	H_0 rejeitada

A Tabela 10 apresenta os resultados dos testes K-S para sequência de pesos das arestas dos grafos de Cenário 1. São apresentados somente o resultado da distribuição que teve o melhor ajuste ao conjunto de dados testado. Para esse conjunto de testes, N é o número de arestas.

Tabela 10 – Resultado dos testes Kolmogorov-Smirnov da sequência de pesos das arestas dos grafos do Cenário 1 com número de órgãos superior a 6.

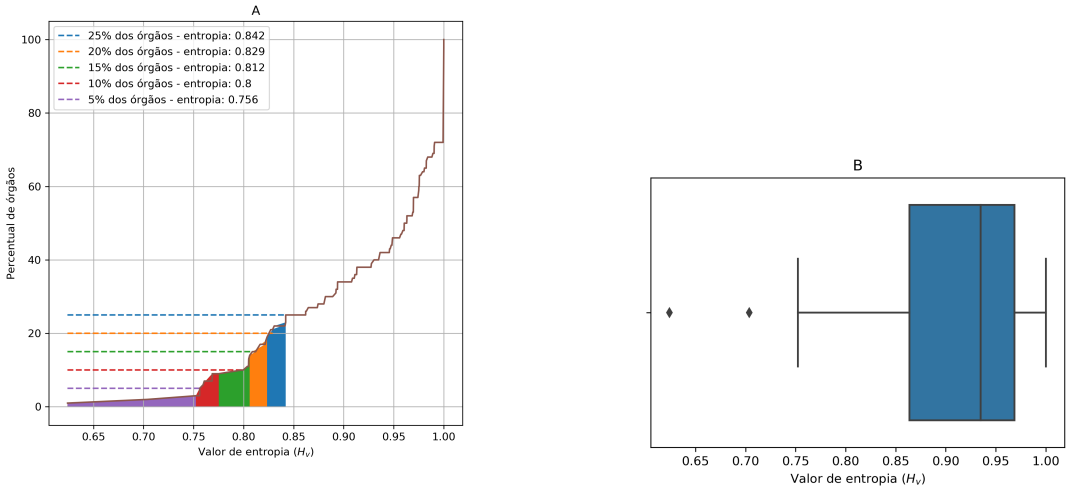
Rede ($ncnae$)	N	$D_{\alpha=0.05}(N)$	Nome da distribuição	D	Resultado do teste
18113	60	0,175	rice	0,393	H_0 rejeitada
18130	94	0,140	levy	0,317	H_0 rejeitada
33121	37	0,218	levy	0,317	H_0 rejeitada
43223	79	0,153	nct	0,386	H_0 rejeitada
43291	38	0,215	f	0,369	H_0 rejeitada
45307	54	0,185	invgamma	0,292	H_0 rejeitada
46397	94	0,140	invgamma	0,171	H_0 rejeitada
46427	37	0,218	lomax	0,294	H_0 rejeitada
46443	580	0,056	halfgennorm	0,136	H_0 rejeitada
46451	1297	0,038	invgamma	0,184	H_0 rejeitada
46460	201	0,096	f	0,288	H_0 rejeitada
46478	92	0,142	invgamma	0,241	H_0 rejeitada
46494	197	0,097	pareto	0,279	H_0 rejeitada
46516	136	0,116	ncx2	0,281	H_0 rejeitada
46648	192	0,098	f	0,298	H_0 rejeitada
47440	24	0,269	lomax	0,299	H_0 rejeitada
47512	70	0,162	alpha	0,321	H_0 rejeitada
47610	40	0,210	f	0,285	H_0 rejeitada
47890	64	0,170	ncx2	0,276	H_0 rejeitada
49230	40	0,210	invweibull	0,368	H_0 rejeitada
49302	82	0,150	rayleigh	0,411	H_0 rejeitada
56201	118	0,125	ncf	0,367	H_0 rejeitada
62015	35	0,224	f	0,461	H_0 rejeitada
62040	32	0,234	invgamma	0,315	H_0 rejeitada
62091	45	0,198	pareto	0,217	H_0 rejeitada
65111	84	0,148	halfgennorm	0,428	H_0 rejeitada
65120	52	0,188	rayleigh	0,433	H_0 rejeitada
69206	64	0,170	halflogistic	0,462	H_0 rejeitada
77110	58	0,178	nct	0,456	H_0 rejeitada
77331	101	0,135	rayleigh	0,428	H_0 rejeitada
77390	95	0,139	maxwell	0,297	H_0 rejeitada
80111	210	0,094	halflogistic	0,462	H_0 rejeitada
80200	53	0,187	powerlognorm	0,490	H_0 rejeitada
81222	106	0,132	invgauss	0,343	H_0 rejeitada
81290	92	0,142	genexpon	0,343	H_0 rejeitada
82300	32	0,234	rayleigh	0,419	H_0 rejeitada
82997	33	0,231	rayleigh	0,394	H_0 rejeitada
95118	55	0,183	invgamma	0,360	H_0 rejeitada

A partir dos resultados do teste K-S para a sequência de graus dos vértices dos grafos G_1 não é possível dizer que nenhum grafo siga uma distribuição específica. Dessa forma, não é possível construir redes sintéticas a partir dos padrões observáveis dos grafos reais do cenário 1. Para gerar os grafos simulados, seria necessário que a sequência de graus dos vértices do conjunto V_1 aderisse a alguma das distribuições estatísticas testadas ou que a sequência de graus dos vértices do conjunto V_o e do conjunto V_e do mesmo grafo G_1 aderisse a alguma daquelas distribuições estatísticas.

5.3.2 Abordagem alternativa

Considerando não ser possível criar novos grafos com base em dados sintéticos e considerando que os órgãos de controle somente realizam fiscalizações em um número inferior ao total de órgãos controlados. Montamos algumas configurações centradas no percentual de órgãos controlados que o TCE-RJ selecionaria para auditar. Conforme discutido na Seção 2.1.4, o percentual de órgão fiscalizados por ano é sempre em número muito inferior a quantidade total de órgãos sob jurisdição. Isso se deve a capacidade operacional dos órgãos de controle. O percentual de órgãos auditados pelo TCE-RJ, que no ano de 2018 foi de 5%, varia ano-a-ano em função de uma série de fatores. A Figura **??** apresenta valores de cortes para o valor de entropia centrada no percentual de órgãos que se pretende auditar. Os órgãos em que em qualquer das redes de G_1 tiveram seu valor de entropia abaixo do valor de corte são selecionadas para a realização de auditoria. O gráfico da Figura **??**(a) apresenta no eixo x os resultados dos valores de entropia dos órgãos. Caso o órgão figure em mais de uma das redes e tenham mais de uma valor de entropia, o valor de entropia adotado é o menor deles. Em função do valor de entropia de x, o eixo y apresenta cumulativamente o percentual de órgãos para aquele valor de entropia. Figura **??**(b) apresenta o boxplot para esta mesma série de valores de entropia do gráfico da Figura **??**(a)

Na primeira configuração, o órgão de controle teria somente a capacidade de auditar 5% dos órgãos sob jurisdição com a metodologia desenvolvida. Esse foi o percentual de órgãos controlados que passaram por auditoria em contratos no ano de 2018. Nessa configuração, o valor de corte para o valor de entropia é 0,756.



(a) (b)

Figura 18 – Valor de corte da entropia centrada no percentual de órgãos controlados a serem auditados

Apesar de 5% ter sido o percentual de auditorias sob licitações e contratos esse não é um percentual que se mantém constante. Como discutido na Seção 2.1.4, o percentual de órgãos a serem auditados por ano oscila em função de uma série de fatores. Tendo isso em vista, montamos outras opções de percentuais. Na segunda configuração, o órgão de controle teria somente a capacidade de auditar 10% dos órgãos sob jurisdição com a metodologia desenvolvida. Nessa configuração, o valor de corte para o valor de entropia é 0,8. Na terceira configuração, o órgão de controle teria somente a capacidade de auditar 15% dos órgãos sob jurisdição com a metodologia desenvolvida. Nessa configuração, o valor de corte para o valor de entropia é 0,812. Na quarta configuração, o órgão de controle teria somente a capacidade de auditar 20% dos órgãos sob jurisdição com a metodologia desenvolvida. Nessa configuração, o valor de corte para o valor de entropia é 0.829. Na quinta configuração, o órgão de controle teria somente a capacidade de auditar 25% dos órgãos sob jurisdição com a metodologia desenvolvida. Nessa configuração, o valor de corte para o valor de entropia é 0.842;

5.4- Resultados dos experimentos computacionais para o Cenário 2

Aplicamos o método desenvolvido na Seção 4.1.2 aos dados descritos na Seção 5.2. Geramos o grafo G_2 usando o Algoritmo 5. Note que, diferentemente do Cenário 1, não há um processo para classificar os lotes de licitações pela CNAE. Há somente um grafo com todos os licitantes independentemente do CNAE do lote da licitação. O grafo G_2 gerado tem 6.390 (seis mil trezentos e noventa) vértices, 116.938 (cento e dezesseis mil novecentos e trinta e oito) arestas, e com densidade dada por $d_{G_2} = 0,00573$. A Figura 19 apresenta o G_2. A seguir, apresentamos as variáveis descritas na Equação 8 que influenciam nos pesos das arestas do grafo G_2 :

$$\sum_{(v_i,v_j)\in E_2} nb_{ij} = 1612718, \qquad fb = 1,$$

$$\sum_{(v_i,v_j)\in E_2} na_{ij} = 70, \qquad fa = \frac{\sum_{(v_i,v_j)\in E_2} nb_{ij}}{\sum_{(v_i,v_j)\in E_2} na_{ij}} = 23038,83,$$

$$\sum_{(v_i,v_j)\in E_2} nr_{ij} = 2372, \qquad fr = \frac{\sum_{(v_i,v_j)\in E_2} nb_{ij}}{\sum_{(v_i,v_j)\in E_2} nr_{ij}} = 679,90$$

$$\sum_{(v_i,v_j)\in E_2} nc_{ij} = 227, \qquad fc = \frac{\sum_{(v_i,v_j)\in E_2} nb_{ij}}{\sum_{((v_i,v_j))\in E_2} nc_{ij}} = 7104,48,$$

$$\sum_{(v_i,v_j)\in E_2} nt_{ij} = 159, \qquad ft = \frac{\sum_{(v_i,v_j)\in E_2} nb_{ij}}{\sum_{(v_i,v_j)\in E_2} nt_{ij}} = 10142,89,$$

$$\sum_{(v_i,v_j)\in E_2} ne_{ij} = 272715, \qquad fe = \frac{\sum_{(v_i,v_j)\in E_2} nb_{ij}}{\sum_{(v_i,v_j)\in E_2} ne_{ij}} = 5,91,$$

$$\sum_{(v_i,v_j)\in E_2} nm_{ij} = 48, \qquad fm = \frac{\sum_{(v_i,v_j)\in E_2} nb_{ij}}{\sum_{(v_i,v_j)\in E_2} nm_{ij}} = 33598,29,$$

$$\sum_{(v_i,v_j)\in E_2} np_{ij} = 807, \qquad fp = \frac{\sum_{(v_i,v_j)\in E_2} nb_{ij}}{\sum_{(v_i,v_j)\in E_2} np_{ij}} = 1998,41$$

Percebe-se que os vínculos entre empresas que têm maior relevância para esse cenário são, pela ordem decrescente: o endereço de e-mail, o endereço físico, o número de telefone, o contador, os sócios, os familiares e os empregados. Essa ordem faz sentido, face ao caso real trazido na Seção 2.1.2 em que esses vínculos revelaram o esquema de

conluio entre aquelas empresas.

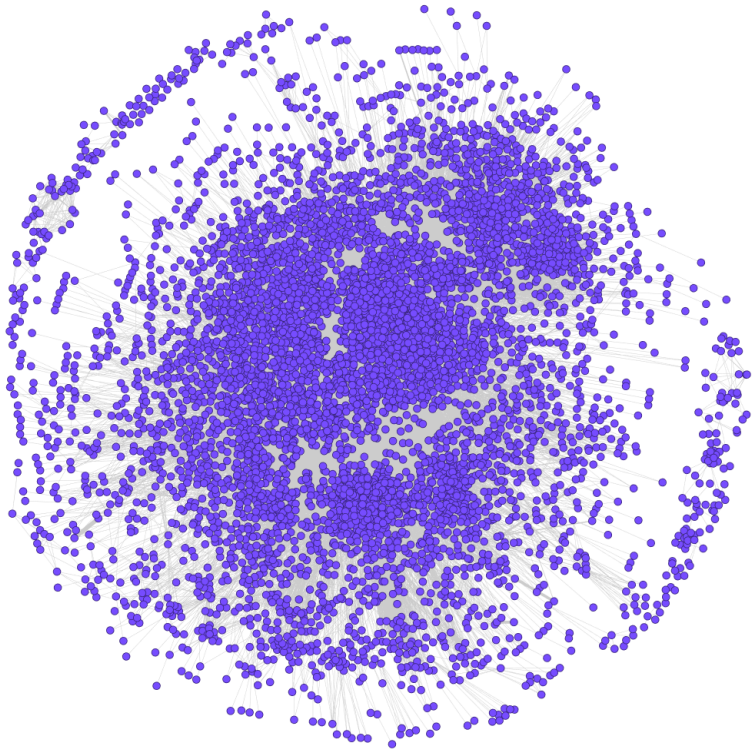


Figura 19 – Grafo G_2

5.4.1 Detecção de comunidades: método de Girvan-Newman

Aplicamos o Algoritmo 6 para identificar comunidades no grafo G_2 e, após aproximadamente 482 horas de processamento, num computador equipado com processador Intel Core I7 4790 3.60ghz de 4ª geração e 8 GB de memória, o o algoritmo teve sua execução interrompida e as comunidades obtidas até este momento foram coletadas. Dado o custo computacional do Algoritmo de GN, cuja complexidade é de $O(|E|^2.|V|)$ e que foi discutida na Seção 2.2.3, o uso do método GN para grafos da ordem de grandeza do grafo G_2 se mostrou inviável para ser executado até o fim. Entretanto, obtivemos resultados parciais com a função $GirvanNewman(G_2)$ que gerou 22 partições p_i para $i = 0, \ldots, 21$. A modularidade e o número de comunidades de cada partição podem ser observados na Tabela 11. Note que a modularidade aumenta à medida em que o algo-

ritmo prossegue dividindo o grafo G_2 em k_i comunidades. A partição p_{21} detectada pelo método GN é a que apresenta a maior modularidade com 83 comunidades detectadas. A implementação do método GN pelo NetworkX, que utilizamos neste trabalho, possui tratamento adequado para o caso de grafos com uma ou mais componentes conexas.

Tabela 11 – Modularidade das partições p_i do grafo G_2

p_i	k_i**(número de comunidades)**	Q **(modularidade)**
p_0	62	0.018487
p_1	63	0.028742
p_2	64	0.028751
p_3	65	0.028775
p_4	66	0.028784
p_5	67	0.028770
p_6	68	0.028773
p_7	69	0.028778
p_8	70	0.028799
p_9	71	0.032510
p_{10}	72	0.032911
p_{11}	73	0.032948
p_{12}	74	0.032960
p_{13}	75	0.034944
p_{14}	76	0.034951
p_{15}	77	0.034953
p_{16}	78	0.034954
p_{17}	79	0.034954
p_{18}	80	0.034955
p_{19}	81	0.034958
p_{20}	82	0.042772
p_{21}	83	0.042773

A Figura 20 apresenta o grafo G_2. Cada comunidade de vértices obtida pelo algoritmo GN é apresentada com uma cor distinta. Verifica-se que uma das comunidades abrangeu a maior parte dos nós do Grafo G_2. Essa comunidade gigante congrega

92,91% dos vértices dessa rede. É comum observar em redes com número elevados de vértices a formação de comunidades que abranjam a maior parte dos vértices. Para o tipo de irregularidade pesquisada neste trabalho e discutido na Seção 2.1.2, em que as empresas formam grupos para agir em conluio de maneira a simular à competitividade em certames de licitações, é comum observar grupos pequenos. Grupos de empresas em conluio formados por muitas empresas tendem a ser instáveis. Entre as linhas 7 e 10 do Algoritmo 6 são calculadas a razão entre o peso e o número das arestas que conectam os vértices de cada comunidade $C_{k_i,i}$ da partição p_{21}. Essa etapa mede a coesão das comunidades identificadas.

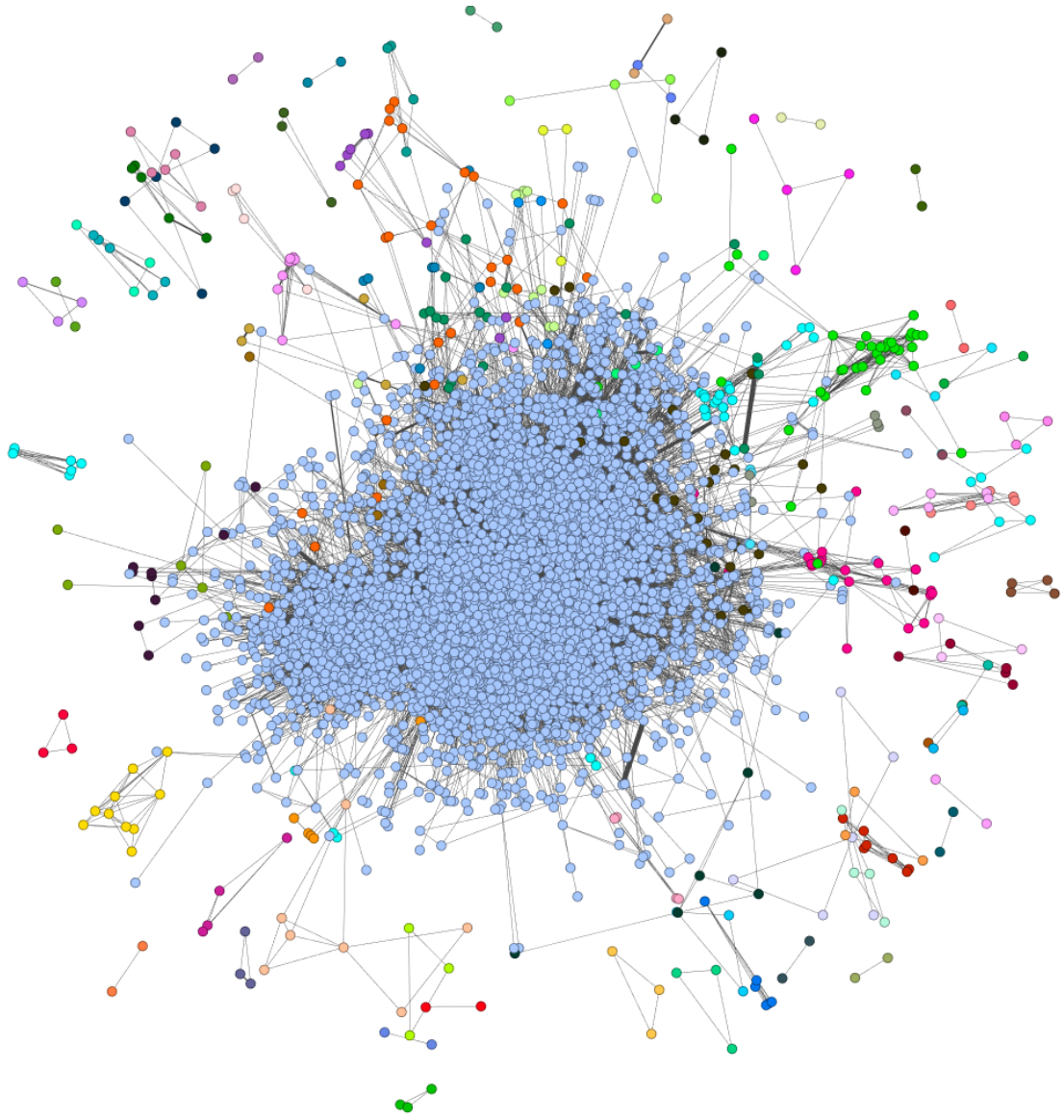


Figura 20 – Grafo G_2 com as comunidades em p_{21}

As 10 (dez) comunidades com maior valor de coesão podem ser observadas na Tabela 12. A coluna "$d_{C_{k_{21}}}$ (densidade)" apresenta a densidade da comunidade identificada que foi apresentada na Seção 2.2 e é denotada pela expressão $d_{C_{k_{21}}} = 2e/v(v-1)$. A coluna "Número de arestas" apresenta o número de arestas que ligam os vértices da comunidade $C_{k_{21}}$. A coluna "Número de vértices" apresenta o número de vértices das comunidade $C_{k_{21}}$. A coluna "Soma dos pesos das arestas" apresenta a soma dos pesos das arestas que ligam os vértices da comunidade $C_{k_{21}}$. A coluna "Razão" apresenta a razão da soma dos pesos das arestas que ligam os vértices da $C_{k_{21}}$ sob o número de arestas que ligam os vértices da comunidade $C_{k_{21}}$. A coluna "Vínculos em $C_{k_{21}}$" apresenta o tipo de vínculo descritos na Seção 4.1.2, encontrado na comunidade $C_{k_{21}}$ e a respectiva quantidade.

Cabe destacar que a densidade igual 1 (um) indica uma comunidade em clique. Ou seja, todo vértice é adjacente a todos os demais vértices dessa comunidade. Entre as comunidades detectadas por esse algoritmo, a comunidade C_{50} apresenta a maior coesão. Essa comunidade tem densidade igual a 1 (um), ela possui 1 (uma) aresta e 2 (dois) vértices. A soma dos pesos das arestas é de 23.042 (vinte e três mil e quarenta e dois), a razão do peso sob o número de arestas é de 23.042 (vinte e três mil e quarenta e dois). As duas empresas dessa comunidade participaram de 4 (quatro) licitações disputando pelo mesmo lote e essas empresas estão localizadas no mesmo endereço. O peso da aresta dessa comunidade se deve a esses vínculos.

Também se destaca a comunidade C_{19} com maior número de vértices em clique dessa relação. Essa comunidade tem densidade igual a 1 (um), ela possui 10 (dez) arestas e 5 (cinco) vértices. A soma dos pesos das arestas é de 6.190 (seis mil cento e noventa), a razão do peso sob o número de arestas é de 619 (seiscentos e dezenove). Nesta comunidade, há 19 (dezenove) vínculos por lotes de licitação, 3 (três) vínculos entre empresas com os mesmos sócios e 30 (trinta) vínculos entre empresas com os mesmos empregados. A Figura 21 apresenta o grafo G_2 somente com as comunidades relacionadas na Tabela 12. As comunidades C_{50} e C_{19} foram identificadas neste Figura.

Todas as 4 (quatro) empresas para as quais se tinha rótulo indicando irregularidade foram agrupadas na comunidade C_5. Essa comunidade tem densidade igual a 0,007, ela possui 14.220.098 arestas e 5.937 vértices ou mais de 92%. A soma dos pesos das arestas é de 14.220.098, a razão do peso sob o número de arestas é de 123. Dada a quantidade de vértices desta comunidade, não seria viável uma análise pormenorizada

em busca de irregularidades dentro desta comunidade.

Figura 21 apresenta o grafo G_2 somente com as comunidades relacionadas na Tabela 12. As comunidades C_{50} e C_{19} foram identificadas neste Figura.

Tabela 12 – Razão dos pesos das arestas sob o número de arestas que conectam os vértices das comunidades em p_{21}

C_i	$d_{C_{k_{21}}}$ (densidade)	Número de arestas	Número de vértices	Soma dos pesos das arestas	Razão	Vínculos em $C_{k_{21}}$
50	1	1	2	23.042	23.042	$nb_{ij}= 4$;$na_{ij}= 1$
33	0,733	11	6	64.492	5.863	$nb_{ij}= 22$;$np_{ij}= 3$; $nt_{ij}= 1$;$nm_{ij}= 1$; $nr_{ij}= 3$;$ne_{ij}= 432$
59	1	3	3	15.284	5.095	$nb_{ij}= 3$;$nc_{ij}= 1$; $nr_{ij}= 12$;$ne_{ij}= 3$
43	1	3	3	9.178	3.059,34	$nb_{ij}= 5$;$nc_{ij}= 1$; $np_{ij}= 1$;$ne_{ij}= 12$
16	0,268	41	18	59.459	1.450	$nb_{ij}= 171$;$np_{ij}= 2$; $nt_{ij}= 1$;$nm_{ij}= 1$; $nr_{ij}= 2$;$ne_{ij}= 9$
34	0,600	6	5	6.005	1.001	$nb_{ij}= 10$;$np_{ij}= 3$
19	1	10	5	6.190	619	$nb_{ij}= 19$;$np_{ij}= 3$; $ne_{ij}= 30$
13	0,246	68	24	16.329	240	$nb_{ij}= 258$;$np_{ij}= 4$; $ne_{ij}= 1.369$
40	0,270	170	36	34.108	201	$nb_{ij}= 537$;$np_{ij}= 9$; $nr_{ij}= 2$;$ne_{ij}= 2.415$
31	1	6	4	844	141	$nb_{ij}= 49$;$ne_{ij}= 135$

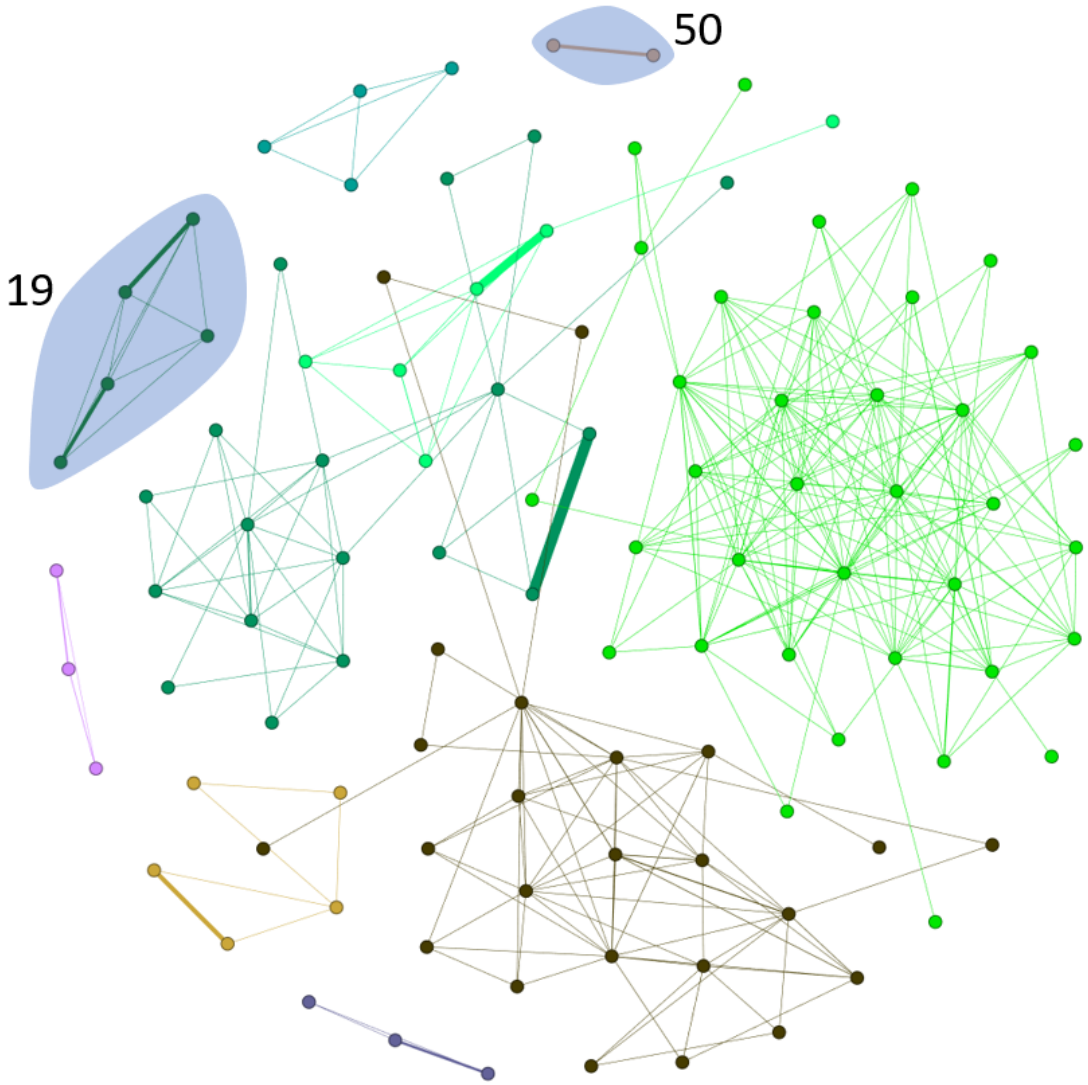


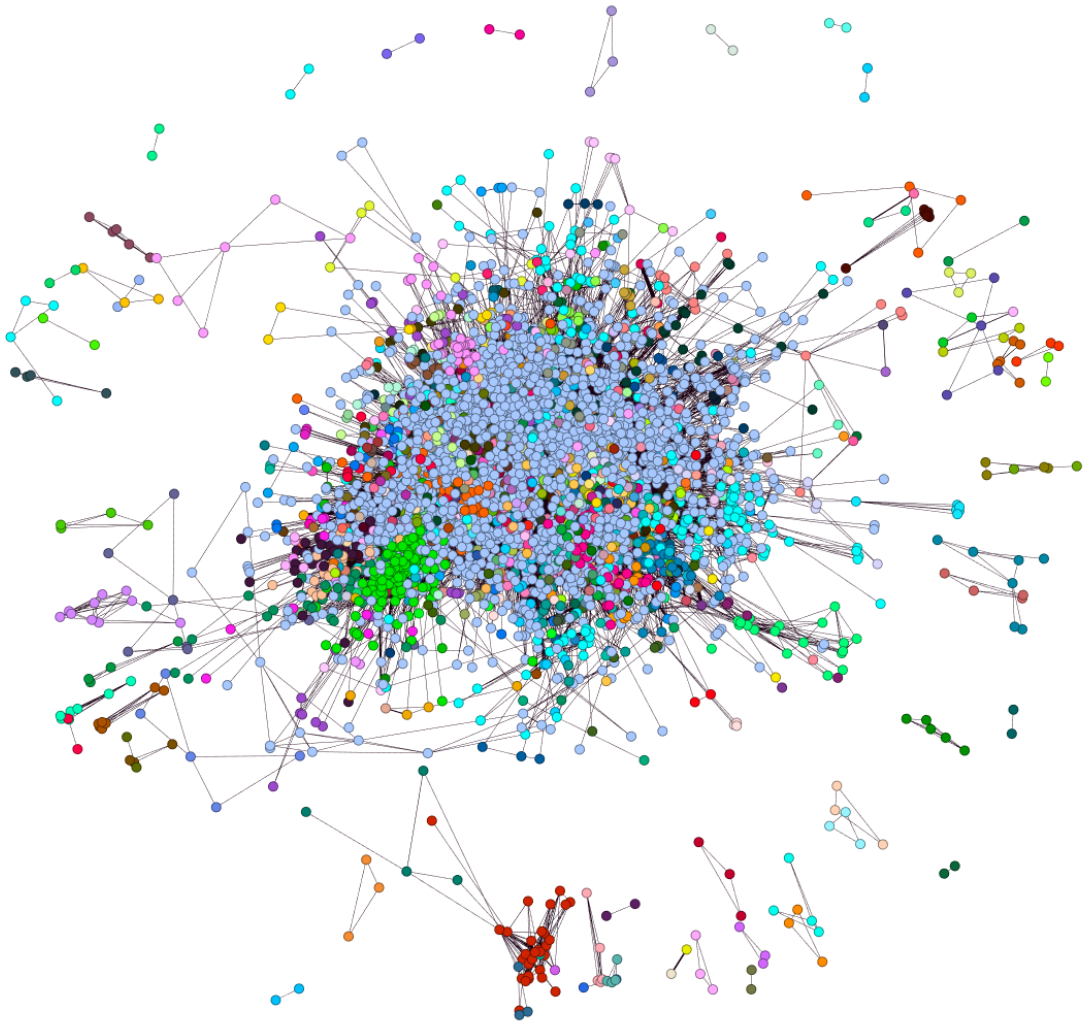
Figura 21 – Grafo G_2 somente com os vértices pertencentes às dez comunidades da Tabela 12

5.4.2 Detecção de comunidades: método do CMN

O Algoritmo 7, da metodologia descrita na Seção 4.1.2 para identificar comunidades no grafo G_2, foi aplicado na base de dados. Cabe destacar que esse algoritmo retorna somente uma partição denominada p_{cnm}. O método CNM identificou 903 comunidades para o grafo G_2. A modularidade dessa partição é de aproximadamente $0,036$. A partição p_{cnm}, resultado do método CNM, tem modularidade menor do que a partição

p_{max}, resultado do método GN. Diferentemente, do método GN que não terminou de processar em tempo hábil, o método CNM terminou o processamento em menos de dez minutos no mesmo equipamento. A Figura 22 apresenta o grafo G_2 com as comunidades da partição em p_{cnm}.

Figura 22 – Grafo G_2 com as comunidades detectadas pelo método CNM



Entre as linhas 2 e 5 do Algoritmo 7, são calculadas a razão entre o peso e o número das arestas que conectam os vértices de cada comunidade da partição p_{cnm}, cujo resultado das 20 (vinte) comunidades com maiores valores podem ser observados na Tabela 13. A coluna "d_{C_i} (densidade)" apresenta a densidade da comunidade identificada que foi apresentada na Seção 2.2 e é denotada pela expressão $d_{C_i} = 2e/v(v-1)$. A coluna "Número de arestas" apresenta o número de arestas que ligam os vértices da comunidade C_i. A coluna "Número de vértices" apresenta o número de vértices das

comunidade C_i. A coluna “Soma dos pesos das arestas” apresenta a soma dos pesos das arestas que ligam os vértices da comunidade C_i. A coluna “Razão” apresenta a razão da soma dos pesos das arestas que ligam os vértices da C_i sob o número de arestas que ligam os vértices da comunidade C_i. A coluna “Vínculos em C_i” apresenta a quantidade por tipo de vínculo descritos na Seção 4.1.2 para cada comunidade C_i.

Entre as comunidades detectadas por esse algoritmo, a comunidade C_{233} apresenta a maior coesão. Essa comunidade tem densidade igual a 1 (um), ela possui 1 (uma) aresta e 2 (dois) vértices. A soma dos pesos das arestas é de 4.045 (quatro mil e quarenta e cinco), a razão do peso sob o número de arestas é de 4.045 (quatro mil e quarenta e cinco). As duas empresas dessa comunidade participaram de 1 (uma) licitação disputando pelo mesmo lote. Essas empresas tem 2 (dois) sócios em comum e 8 (oito) empregados em comum.

Também se destaca a comunidade C_{92} que é a mesma comunidade C_{19} identificada pelo Algoritmo 6, na Seção 4.1.1. Essa comunidade é a comunidade em clique com o maior número de vértices dessa relação. Essa comunidade tem densidade igual a 1, ela possui 10 (dez) arestas e 5 (cinco) vértices. A soma dos pesos das arestas é de 6.190 (seis mil cento e noventa), a razão do peso sob o número de arestas é de 619 (seiscentos e dezenove). Nesta comunidade, há 19 (dezenove) vínculos por lotes de licitação, 3 (três) vínculos entre empresas com os mesmos sócios e 30 (trinta) vínculos entre empresas com os mesmos empregados.

Todas as 4 (quatro) empresas para as quais se tinha rótulo indicando irregularidade foram agrupadas na comunidade C_127. Essa comunidade tem densidade igual a 1, ela possui 6 arestas e 4 vértices. A soma dos pesos das arestas é de 8.215, a razão do peso sob o número de arestas é de 1.369. O método GN manteve esses vértices na maior comunidade da partição. Diferentemente do método GN, o método CNM agrupou essas empresas dentro de uma comunidade com um tamanho em que é possível uma análise pormenorizada por um auditor. O vínculo dessas empresas era esperado, já que o relatório de auditoria que indicou as irregularidades encontradas nas licitações que estas empresas participaram já apontava os vínculos de parentescos dos sócios das empresas e o emprego dos mesmos funcionários.

A Figura 23 apresenta o grafo G_2, mas somente com os vértices das comunidades da Tabela 13 visíveis.

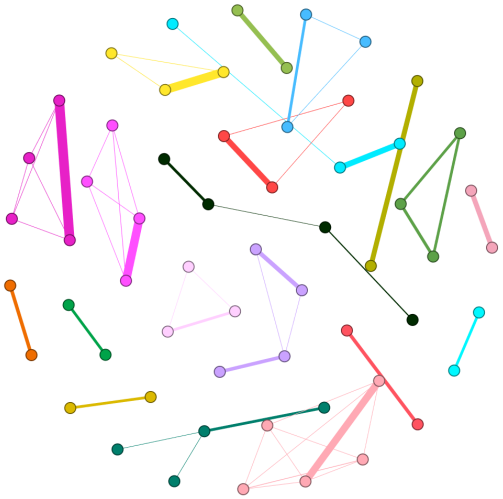


Figura 23 – Grafo G_2 somente com os vértices das comunidades relacionadas na Tabela 13 visíveis

Tabela 13 – Razão dos pesos das arestas sob o o número de arestas das comunidades em p_{cnm}

C_i	d_{C_i} (densidade)	Número de arestas	Número de vértices	Soma dos pesos das arestas	Razão	Vinculos em C_i
233	1	1	2	4045	4.045	$nb_{ij}=1$; $np_{ij}=2$; $ne_{ij}=8$
294	1	1	2	4021	4.021	$nb_{ij}=1$; $np_{ij}=2$; $ne_{ij}=4$
239	1	1	2	3997	3.997	$nb_{ij}=1$; $np_{ij}=2$
267	1	1	2	2732	2.732	$nb_{ij}=1$; $nr_{ij}=4$; $ne_{ij}=2$
246	1	1	2	2727	2.727	$nb_{ij}=8$; $nr_{ij}=4$
147	1	3	3	7149	2.383	$nb_{ij}=21$; $nc_{ij}=1$; $ne_{ij}=4$
170	0,66	2	3	4727	2.364	$nb_{ij}=4$; $np_{ij}=1$; $nr_{ij}=4$; $ne_{ij}=1$
154	1	3	3	6270	2.090	$nb_{ij}=4$; $np_{ij}=3$; $ne_{ij}=46$
254	1	1	2	2032	2.032	$nb_{ij}=10$; $np_{ij}=1$; $ne_{ij}=4$
212	1	1	2	2000	2000	$nb_{ij}=2$; $np_{ij}=1$
280	1	1	2	1999	1999	$nb_{ij}=1$; $np_{ij}=1$
151	1	3	3	4782	1594	$nb_{ij}=5$; $np_{ij}=1$; $nr_{ij}=4$; $ne_{ij}=10$
119	0,66	4	4	6176	1544	$nb_{ij}=22$; $np_{ij}=1$; $nr_{ij}=6$; $ne_{ij}=13$
127	1	6	4	8215	1369,16	$nb_{ij}=9$; $nr_{ij}=12$; $ne_{ij}=8$
135	1	6	4	7234	1205,66	$nb_{ij}=130$; $nc_{ij}=1$
116	0,5	3	4	2607	869	$nb_{ij}=609$; $np_{ij}=1$
125	0,5	3	4	2032	677,33	$nb_{ij}=5$; $np_{ij}=1$; $ne_{ij}=5$
180	1	3	3	2003	667,66	$nb_{ij}=5$; $np_{ij}=1$
171	1	3	3	2001	667	$nb_{ij}=3$; $np_{ij}=1$
92	1	10	5	6190	619	$nb_{ij}=19$; $np_{ij}=3$; $ne_{ij}=30$

Comparando os resultados dos algoritmos GN e CNM, os resultados encontrados tem diferenças importantes, como o tempo de processamento e a quantidade de comunidades descobertas.Quanto a qualidade, o valor da modularidade da partição resultado do método CNM foi de 0,036. O resultado da melhor partição quanto a modularidade do método GN foi de 0,042, a partição p_{21}. Isso é cerca de 16% superior a modularidade da partição resultado do método CNM.

O tempo de execução dos dois algoritmos é a diferença mais marcante. Foram necessários mais de duas semanas de processamento para que o algoritmo GN produzisse resultados parciais com 22 partições. Considerando que o limite do número de partições é o número de vértices do grafo. Considerando que o algoritmo mantivesse a taxa de performance até o final de sua execução, seria necessário pouco mais de 11 anos de processamento para que o algoritmo terminasse o processamento completo. É possível supor que a taxa de processamento por partição aumentasse e o tempo diminuísse, já que a medida que as arestas são removidas, processar a próxima aresta com maior valor de intermediação para removê-la se tornaria uma tarefa mais rápida em grafos esparsos. Contudo, a velocidade de processamento do algoritmo CNM foi de 10 (dez) minutos. Sua Performance foi muito superior ao do algoritmo GN.

Apesar de o método GN ter produzido um resultado com maior qualidade traduzido no valor de modularidade, seu tempo de processamento é impeditivo para processar tantas informações. Cabe destacar que os dados trabalhados nesta pesquisa são oriundos de 84 órgãos. O TCE-RJ possui quase 800 (oitocentos) órgãos jurisdicionados. Uma análise sobre as licitações desse total de órgãos, certamente aumentaria o tamanho dos grafos e por conseguinte o tempo de execução dos algoritmos. O que, na prática, torna impossível a adoção do método GN para o uso pretendido.

Cabe destacar que os auditores selecionam exclusivamente órgãos e seus contratos a serem auditados e não podem abordar diretamente empresas privadas. Tendo em vista a necessidade explicitar os órgãos conexos as comunidades identificadas neste cenário, na próxima Seção, Seção 5.5, propomos uma nova modelagem em grafos reunindo informações a cerca das licitações das empresas, dos órgãos e os resultados das abordagens do Cenário 1 e do Cenário 2.

5.5- Resultados das abordagens dos Cenário 1 e 2

Geramos um novo Grafo G_3. O grafo G_3 é formado pelos vértices de empresas das comunidades com maior coesão da partição p_{21}, Tabela 12, e por órgãos cujas licitações aquelas empresas tenha participado. Além das arestas oriundas do grafo G_2, haverá aresta entre um vértice de empresa e um vértice de órgão quando a empresa tiver participado de uma licitação naquele órgão. O peso dessa aresta é dado pela quantidade de vezes em que uma empresa participou de licitações naquele órgão. O grafo da Figura 24 apresenta o grafo G_3.

Os vértices maiores na cor azul representam o conjunto de órgãos. Os demais vértices representam o conjunto das empresas. Cada cor distinta representa uma comunidade detectada da Tabela 12. O label dos vértices do conjunto de órgãos tem a informação do rótulo do órgão e do menor valor de entropia para aquele órgão dentre todas os Grafos formados no Cenário 1. O tamanho dos vértices do conjunto de órgãos é inversamente proporcional ao seu valor de entropia. Usamos o Yifan Hu, um layout baseado em física, para revelar empresas e órgãos mais fortemente relacionados entre si.

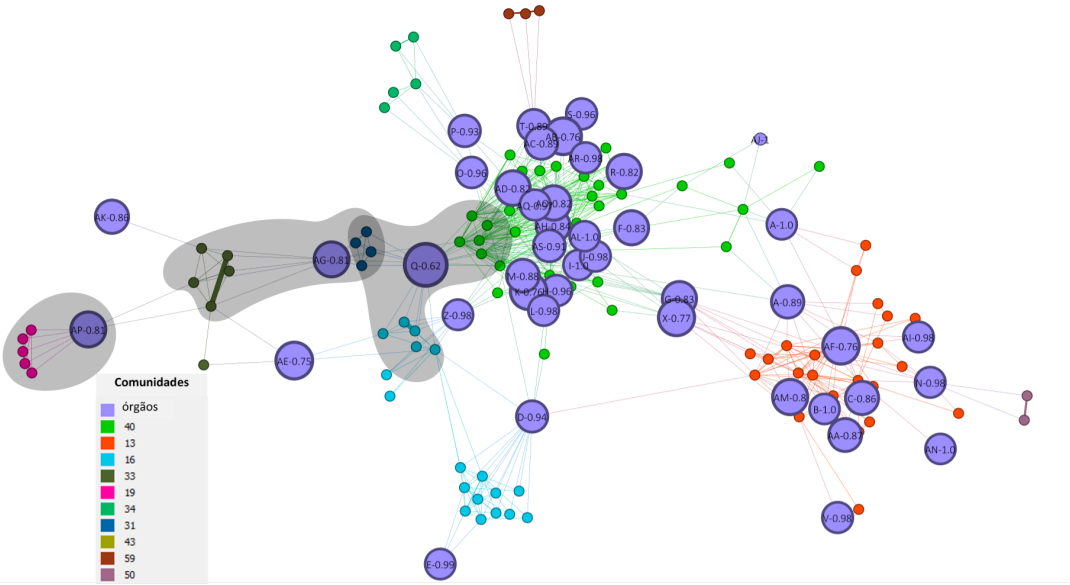


Figura 24 – Grafo G_3 reunindo informações a cerca dos resultados das duas abordagens: os vértices na cor azul representando os valores de entropia e os demais vértices empresas. Cada cor destaca a qual comunidade o vértice pertence

Analisando o grafo formado, podemos destacar alguns órgãos que foram marcados na Figura. O órgão “AP” com valor de entropia de 0.81 e fortemente conexo a todos os vértices de empresas da comunidade 19. Como já tratado anteriormente na Seção 4.1.2, a comunidade 19 é a maior das comunidades em clique detectadas. A comunidade 19 também foi uma das que foi detectadas pelos dois algoritmos, GN e CNM.

O órgão “Q” também merece destaque por ter o menor valor de entropia entre todos os órgãos e estar conexo a três comunidades distintas, as comunidades “16”, “31” e “40”. Esse órgão mantém relação com todos as empresas da comunidade “31”. A comunidade “31” é também uma clique, tem 49 vínculos de participação de licitação e 135 vínculos de empregados.

O órgão “AG”, também merece destaque por estar conexo a todos as empresas da comunidade “31” e a 5 das 6 empresas da comunidade “33”. Das comunidades da Tabela 12, a comunidade “33” é a que apresenta o maior número de vínculos distintos. Além dos 22 vínculos pela participação em licitações, ela também apresenta 3 vínculos de sócios em comum, 1 vínculo pelo mesmo número de telefone, 1 vínculo pelo mesmo endereço de e-mail, 3 vínculos por familiares em comum e 432 vínculos de empregados em comum.

Pelo manual de auditoria e de seleção essa análise preliminar via grafos permite elevar o nível de risco das licitações desses 3 (três) órgãos e já fornece indícios suficientes para justificar verificações adicionais nas suas licitações e contratos, especialmente as licitações e contratações em que essas empresas participam [TCU, 2016; TCE-RJ, 2010].

6- Conclusão

Neste trabalho, apresentamos duas abordagens não supervisionadas para mineração de grafos. Desenvolvemos as duas abordagens e as aplicamos aos dados reais de licitações públicas realizadas por diversos órgãos do estado do Rio de Janeiro entre 2010 e 2018. O objetivo das duas abordagens é identificar irregularidades em licitações públicas. A primeira abordagem é centrada no valor de entropia dos vértices de um grafo bipartido formado por órgãos e empresas licitantes. A segunda abordagem é centrada na detecção de comunidades em um grafo formado exclusivamente por empresas licitantes.

Na primeira abordagem, desenvolvemos uma metodologia que compreendeu as etapas de modelar as licitações públicas, os órgão públicos e as empresas em um grafo bipartido. Calculamos a entropia de cada nó que representa um órgão público e com base nisso propusemos valores de corte de entropia centradas na capacidade operacional de fiscalização do próprio órgão de controle.

Na segunda abordagem, construímos um grafo em que os nós representam empresas licitantes e as arestas conectam empresas que tenham participado da mesma licitação. Os pesos das arestas representam algum vínculo como sócios em comum, mesma sede e mesmo telefone. Usamos os algoritmos GN e CNM para detectar as comunidades no grafo. As comunidades com coesão alta podem indicar empresas que agem em conluio para simular concorrência em licitações públicas.

Um grafo reunindo os resultados das duas abordagens anteriores é produzido e uma indicação de órgãos com indícios de irregularidades e de empresas que provavelmente possam agir em conluio é apresentada para o órgão fiscalizador.

Os resultados alcançados, nas indicações de risco de irregularidades, ainda que necessitem de comprovação, podem ser um importante subsídio no apoio as auditorias de licitações públicas. Como já afirmado anteriormente, os resultados obtidos precisam ser validados para que se verifique a necessidade de se aperfeiçoar os métodos utilizados. Esta validação poderá ser feita mediante a realização de futuras auditorias nos órgãos apontados.

6.1- Analise retrospectiva

Implementamos todos os algoritmos desta pesquisa em Python. Usamos o pacote networkx e seus métodos para criar e manipular os grafos trabalhados. O uso desse pacote nos poupou a implementação de algoritmos corriqueiros, no entanto a performance quanto ao tempo de execução e ao consumo de memória de alguns métodos foi pior do que o esperado. Os métodos de detecção de comunidades tiveram um pior desempenho.

O pacote networkx também não apresenta boas soluções para a visualização dos grafos criados. A geração de layouts em alguns caso tinha o tempo de execução proibitivo para grafos grandes. Esse é o caso do layout usando a função de custo de comprimento de caminho Kamada-Kawai que tinha tempo de execução proibitivo para grafos densos com mais de mil vértices e demasiadamente alto até em pequenos grafos.

Usamos o Gephi para desenhar os grafos trabalhados. O Gephi é um pacote de software de análise e visualização de grafos. Se comparado ao networkx, ele tem diversas opções de layout, permite customizar a visualização dos grafos e apresenta um tempo de execução satisfatório.

6.2- Trabalhos Futuros

Há ações planejadas para continuidade do desenvolvimento do presente trabalho. Entre elas, o aperfeiçoamento da etapa de classificação dos lotes de licitação pelo CNAE. A baixa quantidade de lotes classificados pelo Algoritmo 2, somente 20,68% dos lotes, revela a necessidade de ser aperfeiçoar esse procedimento. Já há uma pesquisa em andamento para usar os dados das notas fiscais eletrônica para classificar os bens e serviços das licitações públicas. Isso vai aperfeiçoar o processo de agrupamento das licitações e permitir uma comparação mais efetiva entre licitações dirigidas ao mesmo setor econômico. A maior parte dos grafos formados no Cenário 1 foram relativamente pequenos. Com um novo procedimento de classificar os lotes de licitação em que não se descartasse tantos dados, talvez fosse possível observar padrões nos grafos do cenário 1. É sabido que distribuições estatísticas como a lei de potências são comuns em grafos do

mundo real. É possível que o tamanho pequeno dos grafos tenha impedido a formação desse padrão.

Usar um grafo dirigido em que os vértices das empresas perdedoras tenham arestas apontando para os vértices que ganharam as licitações pode ter como resultado comunidades com uma coesão maior. Seria necessário alterar a modelagem dos grafos e os métodos de detecção de comunidades.

Referências Bibliográficas

Arief, H. A., Saptawati, G. A. P., and Asnar, Y. D. W. (2016). Fraud detection based-on data mining on indonesian e-procurement system (SPSE). In 2016 International Conference on Data and Software Engineering (ICoDSE), pages 1–6.

Balaniuk, R. (2010). A mineração de dados como apoio ao controle externo. Revista do Tribunal de Contas da União, 117:77–84.

Brasil (1988). Constituição federal da rep´˜blica do brasil.

Brasil (1993). Lei Federal n. 8.666, Lei de Licitações e Contratos Públicos.

Brasil (2001). Lei Federal 10.520, Lei do pregão.

Campos, F. (2008). As práticas de conluio nas licitações públicas à luz da teoria dos jogos. Análise Econômica, 26(50).

Carvalho, R. N., Matsumoto, S., Laskey, K. B., Costa, P. C. G., Ladeira, M., and Santos, L. L. (2013). Probabilistic Ontology and Knowledge Fusion for Procurement Fraud Detection in Brazil. In Bobillo, F., Costa, P. C. G., d'Amato, C., Fanizzi, N., Laskey, K. B., Laskey, K. J., Lukasiewicz, T., Nickles, M., and Pool, M., editors, Uncertainty Reasoning for the Semantic Web II, Lecture Notes in Computer Science, pages 19–40. Springer Berlin Heidelberg.

Carvalho, R. N., Sales, L. J., da Rocha, H. A., and Mendes, G. L. (2014). Using Bayesian Networks to Identify and Prevent Split Purchases in Brazil. In Proceedings of the Eleventh UAI Conference on Bayesian Modeling Applications Workshop - Volume 1218, BMAW'14, pages 70–78, Aachen, Germany, Germany. CEUR-WS.org.

Carvalho, V. A. Restrições à concorrência em contratações públicas: uma preocupação global.

Chakrabarti, D. and Faloutsos, C. (2006). Graph mining: Laws, generators, and algorithms. ACM Comput. Surv., 38(1):2–es.

Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. Physical review E, 70(6):066111.

Cook, D. J. and Holder, L. B. (2006). Mining graph data. John Wiley & Sons.

Davydenko, V. I., Morozov, N. V., and Burmistrov, M. I. (2017). Adaptation of Cluster Analysis Methods in Respect to Vector Space of Social Network Analysis Indicators for Revealing Suspicious Government Contracts. In 2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), pages 57–62.

Domingos, S. L., Carvalho, R. N., Carvalho, R. S., and Ramos, G. N. (2016). Identifying IT purchases anomalies in the brazilian government procurement system using deep learning. In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 722–727.

DPDC, D. d. P. e. D. E. S. d. D. E. M. d. J. (2008). Guia prático para pregoeiros e membros de comissões de licitação.

Erven, G. C. G. V., Carvalho, R. N., Holanda, M. T. d., and Ralha, C. (2017). Graph database: A case study for detecting fraud in acquisition of Brazilian Government. In 2017 12th Iberian Conference on Information Systems and Technologies (CISTI), pages 1–6.

Fortunato, S. (2010). Community detection in graphs. Physics reports, 486(3):75–174.

Fraga, A. A. et al. (2017). Detecção de casos suspeitos de fraudes em licitações realizadas nos municípios da paraíba: uma aplicação de técnicas de mineração de dados. Master's thesis, Universidade Federal da Paraíba.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the national academy of sciences, 99(12):7821–7826.

Goldbarg, M. and Goldbarg, E. (2012). Grafos: Conceitos, algoritmos e aplicações. Elsevier.

Goldschmidt, R., Bezerra, E., and Passos, E. (2015). Data mining: conceitos, técnicas, algoritmos, orientações e aplicações. Rio de Janeiro: Elsevier.

Hand, D. J., Mannila, H., and Smyth, P. (2001). Principles of data mining (adaptive computation and machine learning). MIT Press Cambridge, MA.

IBGE, I. (2007). Introdução a classificação nacional de atividades econômicas - cnae versão 2.0.

IBGE, I. (2017). Contas nacionais trimestrais indicadores de volume e valores correntes.

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

Lu, L. and Zhang, M. (2013). Edge Betweenness Centrality, pages 647–648. Springer New York, New York, NY.

Martins Junior, A. and Braz, M. R. (2010). O controle externo por meio de bases de dados. Revista do Tribunal de Contas da União.

Massey, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. Journal of the American Statistical Association, 46(253):68–78.

Melo, B. M. S. d. S. and Ferreira, M. (2016). Análise de dados no planejamento e na execução de auditorias governamentais. Dados, 12(10):23–03.

Mendroni, M. B. (2009). Aspectos gerais e mecanismos legais. São Paulo: Atlas.

Morettin, P. A. and Bussab, W. O. (2017). Estatística básica. Editora Saraiva.

Newman, M. (2010). Networks: An Introduction. Oxford University Press, Inc., New York, NY, USA.

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

PIETRO, M. S. Z. D. (2009). Direito Administrativo, 22 Edição, São Paulo: Atlas, volume 32. Atlas.

Ralha, C. G. and Silva, C. V. S. (2012). A multi-agent data mining system for cartel detection in brazilian government procurement. Expert Systems with Applications, 39(14):11642 – 11656.

Souza, R. and Pereira, A. (2009). A business intelligence methodology for e-government reverse auctions. In 2009 IEEE Conference on Commerce and Enterprise Computing, pages 82–89.

Speck, B. W. (2008). Tribunais de contas. Corrupçãoensaios e críticas, Editora da Universidade Federal de Minas Gerais, Belo Horizonte, pages 551–558.

Sun, P. G. and Sun, X. (2017). Complete graph model for community detection. Physica A: Statistical Mechanics and its Applications, 471:88–97.

Sundfeld, C. A., Câmara, J., Monteiro, V. C., and Rosilho, A. (2018). O valor das decisões do tribunal de contas da união sobre irregularidades em contratos. Revista Direito GV, 13(3):866–890.

TCE-RJ (2010). Manual de auditoria governamental do tribunal de contas do estado do rio de janeiro.

TCE-RJ (2014). Pedido Lei 12.527.

TCE-RJ (2016). Parecer do Tribunal de Contas do Estado do Rio de Janeiro a cerca das Contas de Governo do Estado do Rio de Janeiro.

TCE-RJ (2019). Ato normativo que dispões sobre o planejamento do tribunal de contas do estado do rio de janeiro.

TCU (2010). Licitações & contratos orientaçõoes e Jurisprudência do TCU. TCU, Tribunal de Contas da Uniao, Brasilia.

TCU (2016). Orientações para seleção de objetos e ações de controle.

Transparency-International (2019). Corruption perceptions.