Submission date: 09/11/2022
Resubmission date: dd/month/year
Camera ready submission date: dd/month/year

1st round notification: 11/05/2023
Available online: 31/12/2023
Publication date: 31/12/2023

Section: extended version
Extended version of selected papers from the XI Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2022)

Identificação de licitações suspeitas de fraude por meio de trilhas de auditagem

Identification of suspected fraud bids through audit trails

Lucas L. Costa[1], **Clara A. Bacha**[1], **Gabriel P. Oliveira**[1], **Mariana O. Silva**[1], **Matheus C. Teixeira**[1], **Michele A. Brandão**[1,2], **Anisio Lacerda**[1], **Gisele L. Pappa**[1]

[1]Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG, Brasil

[2]Instituto Federal de Minas Gerais (IFMG) – Ribeirão das Neves, MG, Brasil

{lucas-lage,clarabacha}@ufmg.br
{gabrielpoliveira,mariana.santos,matheus.candido}@dcc.ufmg.br
michele.brandao@ifmg.edu.br, {anisio,glpappa}@dcc.ufmg.br

Abstract. *Different information technologies help to promote government transparency, made possible by agreements promoting and encouraging open data. Public bids are a specific type of this data, made available by the Brazilian government, and aim to ensure transparency and free competition between bidders. However, auditing for irregularities is a non-trivial task due to the massive volume of data and the reduced number of specialists. Thus, this work proposes a methodology based on concepts of audit trails and social networks to create fraud alerts in bids. We also propose an approach to ranking bids according to these tracks. The results reveal that our proposal helps in the fight against corruption by being able to identify suspicious bids.*

Keywords. *Bidding fraud, Social network analysis, Audit trails*

Resumo. *Diferentes tecnologias da informação têm sido utilizadas para promover a transparência governamental, a qual é possibilitada por acordos que promovem e incentivam a abertura de dados. Licitações públicas são um tipo específico desses dados, as quais são disponibilizadas pelo governo brasileiro e visam garantir a transparência, bem como a livre concorrência entre licitantes. Entretanto, a auditoria em busca de irregularidades é uma tarefa não trivial devido ao enorme volume de dados e uma quantidade reduzida de especialistas. Assim, este trabalho propõe uma metodologia baseada em conceitos de trilhas de auditagem e redes sociais para criar alertas de fraude em licitações. Também é proposta uma abordagem para ranquear as licitações de acordo com essas trilhas. Os resultados revelam que nossa proposta auxilia no combate à corrupção por conseguir identificar licitações suspeitas.*

Palavras-Chave. *Fraude em licitações, Análise de redes sociais, Trilhas de auditagem*

Cite as: Costa, L. L., Bacha, C. A., Oliveira, G. P., Silva, M. O., Teixeira, M. C., Brandão, M. A., Lacerda, A. & Pappa, G. L.. (2023). Identificação de licitações suspeitas de fraude por meio de trilhas de auditagem. iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), Vol(16), 13:1-13:23. DOI: 10.5753/isys.2023.3013

1. Introdução

Dados hoje representam recursos valiosos para organizações públicas e privadas por possibilitarem a extração de informações capazes de ajudar na tomada de decisão em diversas esferas [Aquino Jr et al. 2019]. No setor público, vários países firmaram acordo com o intuito de incentivar e fortalecer a transparência governamental. Essa política de abertura dos dados é sustentada pelas Tecnologias de Informação e Comunicação (TICs), as quais também auxiliam no combate à corrupção [Park and Kim 2020].

Dentre os dados abertos brasileiros, encontramos as licitações públicas, que são definidas como um conjunto de procedimentos por meio do qual a Administração Pública realiza compras ou contratações de produtos e serviços. Em geral, uma licitação pública possui três objetivos principais: selecionar a proposta mais vantajosa para a Administração Pública; garantir igualdade de condições a todos que queiram contratar com o Poder Público; e promover o desenvolvimento nacional sustentável.[1]

Apesar da legislação impor mecanismos para impedir fraudes em licitações públicas, escândalos de corrupção ainda são frequentes no Brasil. A checagem dessas licitações é realizada, em geral, manualmente por meio de um monitoramento em tempo real. Em outras palavras, todo o processo licitatório é acompanhado — desde a publicação do edital até que seja firmado e executado o contrato, e verificado o cumprimento do objeto licitado — por pessoas. Essa inspeção manual requer tempo e grande mobilização de recursos humanos. Assim, a combinação de perícia humana com sistemas computacionais é essencial.

Além disso, a detecção de fraudes em licitações é um processo complexo, uma vez que essas fraudes não são pontuais e isoladas, mas envolvem interações diretas, indiretas e até temporais entre as entidades envolvidas. Dentre essas entidades podemos citar os licitantes, representados por fornecedores, sejam eles pessoas físicas ou jurídicas, interessados em fornecer o objeto da licitação ao licitador, ou seja, um órgão público. Portanto, formas automáticas de monitoramento, análise e cruzamento dos dados à procura de infrações são necessárias para auxiliar na identificação de fraudes.

O principal objetivo deste trabalho é desenvolver uma metodologia capaz de identificar indícios de fraudes em licitações públicas por meio da combinação de conceitos de trilhas de auditagem e redes sociais. Esses conceitos são explorados por meio de tecnologias da informação, que facilitam a obtenção de informação e conhecimento sobre fraudes em licitações pelos responsáveis em realizar tal análise.

Para alcançar esse objetivo, a partir de um conjunto de dados de licitações públicas municipais e estaduais de Minas Gerais, um conjunto de trilhas de auditagem foi definido e implementado em parceria com especialistas do Ministério Público de Minas Gerais (MPMG). Uma trilha de auditagem é uma sequência de passos seguidos para identificar indícios de tipos específicos de irregularidades encontrados em fraudes de licitações. Essas trilhas ajudam a selecionar os dados de interesse das bases de dados de licitações. Exemplos simples de trilhas incluem a identificação de licitações contendo um único licitante ou licitantes que possuem sócios em comum.

[1] Lei nº 8.666/93: `http://www.planalto.gov.br/ccivil_03/Leis/L8666cons.htm`

Em seguida, esses dados são pré-processados e modelados como uma rede, que captura a interação entre os licitantes. Com a modelagem na forma de grafo, o arcabouço da análise de grafos pode ser empregado para a verificação de indícios de fraudes. Este artigo amplia os resultados publicados em [Costa et al. 2022], que apresentou dez trilhas de auditagem definidas a partir de uma modelagem de empresas, licitações e pessoas sócias como uma rede social. Os resultados também destacaram que é possível utilizar a metodologia proposta para reduzir o volume de dados a serem analisados por especialistas.

Neste artigo, complementamos nossas análises com a inclusão de duas novas trilhas de auditagem ao modelo e propomos um modelo de ranqueamento de licitações de acordo com o número de alertas de fraude. Ou seja, a partir dos resultados de cada trilha de auditagem, realizamos um ranqueamento das licitações com maior risco de terem sido fraudadas. É importante destacar que, além de reduzir o volume de dados a serem analisados por especialistas e indicar uma lista de prioridade de licitações a serem investigadas, a metodologia também permite gerar subsídios para elaboração de algoritmos capazes de classificar uma licitação como fraude ou não de forma automática. Em geral, as novas avaliações mostram que a metodologia das trilhas de auditagem proposta e o ranqueamento das licitações são um avanço em direção à tarefa de classificação de uma licitação como fraudulenta.

Este trabalho está organizado da seguinte maneira. Os trabalhos relacionados são apresentados na Seção 2 e a base de dados utilizada no trabalho é descrita na Seção 3. A Seção 4 apresenta a metodologia adotada, descrevendo o processo de seleção e pré-processamento dos dados e modelagem do grafo. Já na Seção 5, são apresentadas várias caracterizações do grafo que podem ser usados como indicadores de ilicitude. Na Seção 6, é apresentado um estudo de caso de uma licitação fraudulenta. Em seguida, a Seção 7 descreve uma nova abordagem para ranquear as licitações com base nos alertas gerados pelas trilhas de auditagem. Finalmente, na Seção 8 são abordadas as limitações e desafios deste trabalho e na Seção 9 são apresentadas as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

A modelagem de situações que envolvem indivíduos/organizações e seus relacionamentos como uma rede social possibilita a obtenção de soluções viáveis para resolução de problemas em diferentes contextos. Por exemplo, Florentino et al. [2022] analisam as interações entre conteúdos textuais postados no YouTube para identificar suspeitos de crimes de pedofilia. Já em Kansaon et al. [2019], são utilizados dados de redes sociais e algoritmos de classificação para detectar emoções em postagens de usuários.

No contexto de detecção de fraudes, as redes sociais têm sido utilizadas para obtenção de características sobre colaborações entre suspeitos de cometer fraude [Óskarsdóttir et al. 2022]. Em particular, Óskarsdóttir et al. [2022] identificam fraudes em solicitações de seguros utilizando um ranqueamento baseado em características extraídas de uma rede social construída a partir de dados de solicitações de seguros de uma empresa seguradora. Já em Brum el al. [2022], foi proposta uma abordagem baseada em redes de menções para identificar perfis de bots em redes sociais e mostraram como esses perfis manipulam mensagens no contexto político, em um tipo de fraude mais indireta.

Complementarmente, Vlasselaer et al. [2015] apresentam uma nova abordagem que usa inferência ativa através de um algoritmo baseado em rede para melhor detectar fraudes em redes sociais que variam no tempo. Em outro estudo, os autores introduzem uma nova abordagem que define e extrai recursos de uma rede ponderada no tempo, explorando e integrando recursos intrínsecos na detecção de fraudes [Vlasselaer et al. 2017].

Em conjunto com abordagens de redes sociais, também é importante considerar o uso de técnicas de inteligência artificial para identificar fraudes. Nesse contexto, muitas abordagens propostas para detecção automática de fraudes têm utilizado técnicas da área de mineração de dados e aprendizado de máquina [Elshaar and Sadaoui 2020, Ganguly and Sadaoui 2018, Oliveira et al. 2022a]. Por exemplo, Velasco et al. [2021] propõem uma ferramenta que incorpora algoritmos de mineração de dados para quantificar padrões de risco de corrupção, levando a melhorias na qualidade dos gastos públicos e à identificação de casos de fraude. Com uma abordagem baseada em um algoritmo de aprendizado de máquina supervisionado, Anowar e Sadaoui [2019] desenvolveram um classificador de fraude eficiente que permite distinguir entre licitantes legítimos e não-legítimos. Além de classificadores clássicos, alguns estudos exploraram técnicas mais robustas baseadas em redes neurais para tratar a tarefa de detecção de fraudes [Abidi et al. 2021, Pereira and Murai 2021].

Uma outra abordagem é a mineração de processos, que consiste na combinação entre mineração de dados e análise de processos [Santoro et al. 2020]. Na auditoria de fraudes, Santoro et al. [2020] analisam o benefício do uso de técnicas de mineração de processos em uma aplicação real que consiste em pedidos de pagamento da União Europeia a agricultores alemães do Fundo Europeu de Garantia Agrícola. Esse estudo identificou que apesar de técnicas de mineração de processos auxiliar na auditoria de fraudes, a análise manual das transações ainda é necessária.

Um aspecto importante a ser considerado é que a maioria das abordagens citadas anteriormente necessitam de dados rotulados para o treinamento de modelos de aprendizado ou outras técnicas de mineração de dados. Como alternativas mais simples e eficientes, abordagens baseadas em redes sociais vêm sendo consideradas como boas opções para a tarefa de detecção de fraudes. Por exemplo, Araújo et al. [2021] propõem uma observação detalhada do fluxo de processos judiciais utilizando redes complexas.

Independente da abordagem utilizada, ainda são poucos os estudos tratando especificamente da análise ou detecção de fraude em licitações públicas [Ralha and Silva 2012, Grace et al. 2016, Andrade et al. 2020, Lima et al. 2020], provavelmente, devido à falta de dados adequados para apoiar tal tarefa. No entanto, tais trabalhos não utilizam abordagens baseadas em rede em suas análises. Nesse contexto, este trabalho busca preencher as lacunas de pesquisa existentes ao analisar alertas de fraudes em licitações no estado (e nos municípios) de Minas Gerais através da análise da rede social construída a partir de trilhas de auditagem.

3. Base de Dados

Neste trabalho, foram utilizados os dados fornecidos pelo Ministério Público de Minas Gerais (MPMG) por meio do Programa de Capacidades Analíticas. Esses dados são si-

gilosos e, por isso, pouco detalhados. Os dados foram disponibilizados em um Sistema de Gerenciamento de Bancos de Dados (SGBD), que possui informações de licitações municipais e estaduais, bem como de seus licitantes e sócios.

Para as licitações, foram utilizados dados públicos. As municipais são provenientes do Sistema Informatizado de Contas dos Municípios (SICOM) do TCE-MG,[2] enquanto as estaduais vieram dos dados do Portal da Transparência do TCE-MG.[3,4] Também foi considerada a base de dados do CEIS (Cadastro de Empresas Inidôneas e Suspensas),[5] que apresenta empresas com sanções vigentes em todo o país. Além dos dados públicos, foram utilizados dados sigilosos advindos do SISAP (Sistema Integrado de Administração de Pessoal) e do IPSEMG (Instituto de Previdência dos Servidores do Estado de Minas Gerais) para obter informações sobre servidores públicos e seus dependentes. Por fim, também são considerados dados do SERPRO (Serviço Federal de Processamento de Dados) para enriquecimento das informações das empresas licitantes.

Ao todo, foram consideradas 14.565 licitações estaduais e 363.572 licitações municipais. As licitações estão divididas entre 17 modalidades, incluindo concorrência, convite, concurso, pregão, entre outras. Por fim, nosso conjunto de dados possui informações de 103.858 empresas licitantes distintas que participaram de tais processos licitatórios. Os dados incluem informações relevantes para detecção de vínculo em comum, incluindo endereço, telefones, e-mail e sócios. Parte desses dados foram publicados em [Silva et al. 2022] e uma análise da qualidade desses dados foi apresentada em [Oliveira et al. 2022b].

4. Metodologia para Identificação de Fraudes

Esta seção apresenta a metodologia para identificação de possíveis fraudes em licitações. As principais etapas dessa metodologia são baseadas na construção de trilhas de auditagem, cujo conceito foi discutido na Seção 1. Tais etapas são apresentadas na Figura 1 e incluem: a filtragem dos dados (Seção 4.1), o pré-processamento dos dados filtrados (Seção 4.2), a modelagem de uma rede social a partir de tais dados (Seção 4.3) e a análise dessa rede para levantamento de alertas em fraudes de licitação (Seção 4.4).

É importante destacar que as etapas da metodologia foram realizadas tendo como base a definição de doze trilhas de auditagem, apresentadas na Tabela 1. As primeiras sete trilhas tiveram como objetivo investigar se há alguma irregularidade na licitação ao considerar algumas características do licitante (ex: licitante com CNPJ inativo), sendo portanto denominadas *Trilhas de Licitante (Nós)*. A oitava trilha verifica se há alguma irregularidade relacionada a alguma característica de pessoas sócias dos licitantes, sendo portanto denominada *Trilha de Sócio (Nós)*. Já as quatro trilhas restantes investigam se há irregularidades considerando os vínculos entre os licitantes, sendo portanto *Trilhas de Vínculo (Arestas)*.

[2] http://dadosabertos.tce.mg.gov.br/

[3] https://www.transparencia.mg.gov.br/compras-e-patrimonio/compras-e-contratos

[4] https://fiscalizandocomtce.tce.mg.gov.br/#/inicio

[5] https://www.portaldatransparencia.gov.br/sancoes/ceis

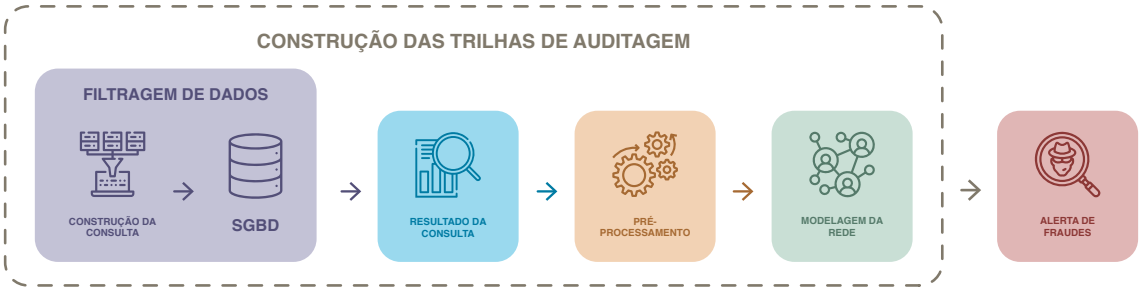

Figura 1. Metodologia para alerta de fraudes em licitações.

Tabela 1. Definição das trilhas de auditagem a serem aplicadas nas licitações.

#	Trilha de Auditagem	Regra
	Trilhas de Licitante (Nós)	
T_1	Licitante licitando antes de registro	Verificar as licitações que contenham participantes que estão licitando antes da empresa iniciar suas atividades
T_2	Licitante licitando com sanção ativa	Verificar as licitações que contenham licitantes com alguma sanção ativa na base do CEIS (Cadastro de Empresas Inidôneas e Suspensas)
T_3	Licitante com CNPJ inativo	Verificar as licitações com a presença de licitantes com o CNPJ inativo
T_4	Licitante perdedor frequente	Identificar licitações com a presença de licitantes com alto percentual de derrotas
T_5	Licitante vencedor frequente	Identificar licitações com a presença de licitantes com alto percentual de vitórias
T_6	Licitante único	Identificar licitações com licitante único
T_7	Licitante com CNAE incongruente	Identificar licitações com licitantes cujo código CNAE (Classificação Nacional de Atividades Econômicas) é incongruente com a descrição dos itens licitados [Oliveira et al. 2022a]
	Trilha de Sócio (Nós)	
T_8	Licitante cujos sócios são ou têm vínculo com servidores públicos	Identificar as licitações que apresentam licitantes cujos sócios são servidores públicos ou dependentes de servidores da entidade que realiza a licitação
	Trilhas de Vínculo (Arestas)	
T_9	Licitantes com sócios em comum	Identificar as licitações com licitantes distintos que possuem pelo menos um sócio em comum
T_{10}	Licitantes com e-mails em comum	Verificar se a licitação possui licitantes distintos com e-mails em comum
T_{11}	Licitantes com telefones em comum	Verificar se a licitação possui licitantes distintos com número de telefone em comum
T_{12}	Licitantes com endereços em comum	Identificar as licitações com licitantes distintos com endereço em comum

4.1. Filtragem dos Dados

Para realizar a tarefa de identificação de possíveis fraudes em licitações, alguns filtros foram criados. Esses filtros consideraram as trilhas de auditagem descritas na Tabela 1, bem como as entidades de interesse envolvidas.

Conforme descrito na Seção 3, os dados de licitações utilizados são públicos, e dessa forma os licitantes que são pessoas físicas tiveram seus CPFs mascarados. Isso impossibilita qualquer processamento sobre esses dados e, portanto, eles são desconsiderados das análises. Sendo assim, são considerados três tipos de entidades de interesse: licitações, licitantes do tipo pessoa jurídica (Empresas) e seus sócios (Pessoas).

Após a análise dos dados disponíveis e das trilhas de interesse, foram elaboradas consultas para obter informações sobre licitações, empresas licitantes e seus sócios de

forma que possam auxiliar na identificação de possíveis fraudes em licitações. Tais consultas retornam dados em formato ainda bruto, o que dificulta a identificação de indícios de fraudes relacionadas às características dos licitantes e das licitações. Dessa forma, são necessárias etapas adicionais de pré-processamento dos dados, descritas a seguir.

4.2. Pré-processamento dos Dados

Considerando as trilhas de auditagem de interesse, após a execução da etapa de filtragem de dados e a partir dos dados retornados pelas consultas, dois tipos de pré-processamento foram executados: padronização de dados e construção de atributos referente às entidades.

Em relação à padronização, foram considerados os campos e-email, endereço e telefone, que fazem parte da entidade licitante. Para o campo de e-mail, verificou-se se o formato era válido com expressões regulares e os valores foram convertidos para letras minúsculas. Para a padronização do campo de endereço, utilizou-se a função *soundex* do SQL (*Structured Query Language*), que converte uma *string* em um código de quatro caracteres baseando-se no som da cadeia de caracteres no idioma inglês. Já para os números de telefone, removeu-se a máscara e a formatação dos valores. Em seguida, foram retirados todos os caracteres não numéricos. Por fim, verificou-se o tamanho do número de telefone resultante, sendo considerado válido um tamanho entre 9 e 11 caracteres.

O segundo tipo de pré-processamento (i.e., construção dos atributos de entidades) é necessário devido às múltiplas fontes de dados utilizadas para construir o conjunto final de dados (Seção 3). Por exemplo, para licitações municipais, as informações de todos os licitantes participantes estão presentes no conjunto de dados. No entanto, as licitações estaduais só possuem informações dos licitantes vencedores. Por isso, para tais licitações, apenas os licitantes vencedores são considerados.

Além disso, para auxiliar na construção das trilhas, foi estabelecida uma data de referência para cada licitação. Para as licitações municipais, essa data corresponde à data de habilitação do licitante no processo licitatório. Nas licitações estaduais, como não havia nenhum campo de data preenchido, foi utilizada a data corrente (i.e., data de execução da trilha) como referência. Dessa forma, a data de referência será sempre mais atual do que as datas que estão sendo comparadas (e.g., data de registro do licitante). Assim, os licitantes não serão prejudicados nas trilhas em caso de ausência de informações.

4.3. Modelagem da Rede Social

A modelagem do problema de geração de alertas de fraudes em licitações foi feita a partir de uma rede social representada por um grafo G, que possui dois tipos de nós: as empresas licitantes (V) e os sócios dessas empresas ($\overline{V}$). Há dois tipos de arestas: entre nós do tipo empresa que participaram de uma mesma licitação (E); e entre nós empresa e pessoa ($\overline{E}$), onde a pessoa é uma sócia da empresa. No caso de arestas entre nós empresa, uma função $f(m, n)$ determina o peso da aresta entre duas empresas m e n de acordo com a quantidade de vínculos que uma empresa licitante tem com outra. Os vínculos levados em conta foram sócios, e-mail, telefones ou endereço em comum, i.e., o domínio de $f(m, n)$ está no intervalo $[1, 4]$. No caso de relações entre empresas e pessoas, a relação só existe quando a pessoa é sócia da empresa. Formalmente:

$G = (\{V \cup \overline{V}\}, \{E \cup \overline{E}\})$, onde

$V = \{v \mid v \in$ *Empresas*$\}$, $\overline{V} = \{\overline{v} \mid \overline{v} \in$ *Pessoas*$\}$,

$E = \big\{(m,n) \mid ((m,n) \in V^2) \Leftrightarrow f(m,n) > 0\big\}$,

$f(m,n) =$ número de vínculos em comum de m com n, onde

vínculo $\in \{$*sócio, endereço, telefone, email*$\}$

$\overline{E} = \big\{(m,\overline{o}) \mid (m \in V \wedge \overline{o} \in \overline{V}) \Leftrightarrow \overline{o}$ é sócio de $m\big\}$

Além disso, os nós da rede social proposta apresentam atributos, utilizados para construir as arestas do grafo e definir as regras que precisam ser seguidas pelas trilhas de auditagem (que serão apresentadas na Seção 4.4). Para o licitante (nó do tipo empresa), temos a data de registro da empresa (*data_registro*), CNPJ, *status* da empresa na Receita Federal (i.e., ativa ou não), *endereço*, *e-mail*, *telefone*, *sócios*, *sanções* e uma lista de códigos CNAE incongruentes com itens licitados (*CNAE_incongruente*). Esta lista foi montada utilizando o arcabouço *Primeiro Termo* [Oliveira et al. 2022a], que combina técnicas de padronização textual e estruturação de atributos para que os itens das licitações, armazenados em formato de texto livre, sejam representados com o mesmo identificador (i.e., *token*). Para o sócio (nó pessoa), temos o CPF, o tipo de *vínculo* com a empresa (e.g., sócio, representante legal), um marcador para indicar se ele é servidor publico (*servidor*) e uma lista de seus parentes que são servidores públicos (*parente_servidor*).

Dado o grafo G, uma licitação é representada pelo grafo $\hat{G}_i = (\{\hat{V}_i \cup \hat{\overline{V}}_i\}, \{\hat{E}_i \cup \hat{\overline{E}}_i\})$, onde $\hat{G}_i \subset G$ e $\hat{V}_i = \{v \mid v \in$ Empresas que participaram da licitação $i\}$, $\hat{\overline{V}}_i = \{\overline{v} \mid \overline{v} \in$ Pessoas sócias das empresas que participaram da licitação $i\}$. As arestas que refletem os vínculos entre as empresas licitantes de $\hat{G}_i$ são representadas pelo conjunto $\hat{E}_i = \{(m,n) \in$ Arestas entre empresas participantes da licitação $i\}$. Já as arestas empresa-pessoa da licitação $\hat{G}_i$ são representadas pelo conjunto $\hat{\overline{E}}_i = \Big\{(m,\overline{o}) \in$ Arestas entre empresas da licitação $\hat{G}_i$ e pessoas sócias dessas empresas$\Big\}$. A Figura 2 ilustra e exemplifica a modelagem proposta com três empresas (v_1, v_2 e v_3) que participam de dois processos licitatórios ($\hat{G}_1$ e $\hat{G}_2$).

Além disso, temos a data de cada licitação e construímos dois novos atributos: a lista de licitantes perdedores e vencedores frequentes. Para a identificação de empresas vencedoras frequentes, foram analisadas empresas que, entre os anos de 2014 e 2021, somaram um valor total homologado maior ou igual a R$400.000,00 (quatrocentos mil reais). Essas empresas também deveriam ter participado de pelo menos dois processos licitatórios com mais de uma empresa participante nas modalidades pregão eletrônico, pregão presencial ou dispensa de licitação por valor. Para uma empresa ser considerada vencedora de uma licitação, é necessário que ela tenha vencido pelo menos 70% do valor total desta, sendo considerada uma vencedora frequente quando houver uma taxa de vitórias de pelo menos 70% das licitações participadas.

Da mesma forma, para a identificação de empresas perdedoras frequentes, foram analisadas empresas que, entre os anos de 2014 e 2021, tenham perdido pelo menos 70% das licitações com uma participação mínima de pelo menos seis licitações por ano. As

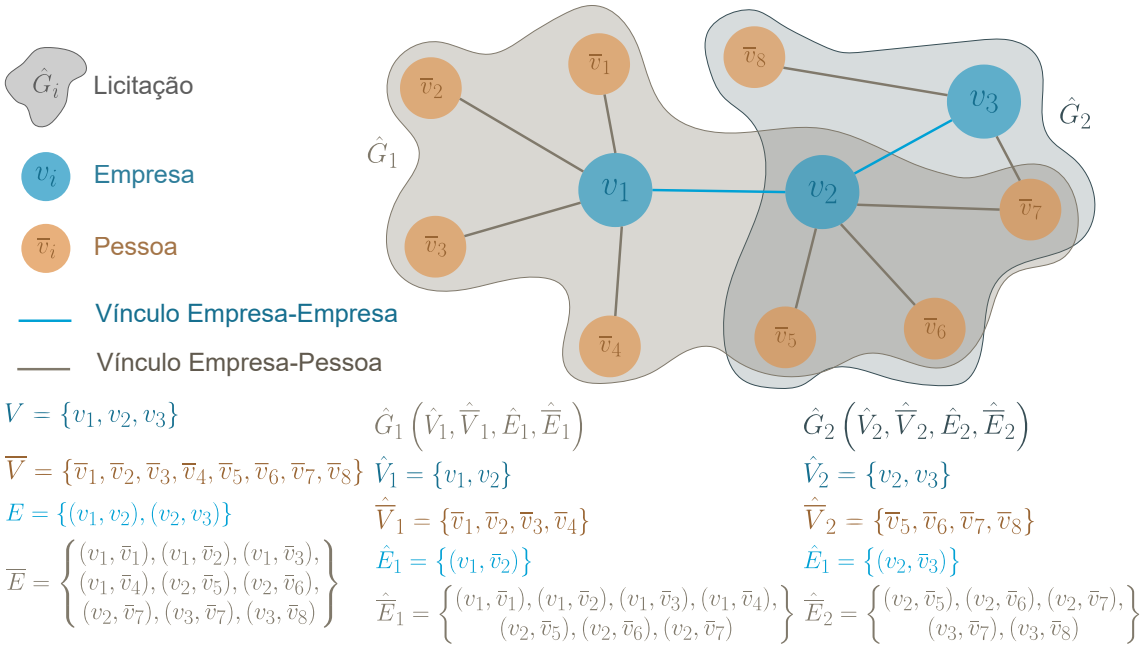

Figura 2. Rede social heterogênea entre empresas licitantes e seus sócios.

modalidades dessas licitações são pregão eletrônico, pregão presencial ou dispensa de licitação por valor, resultando em um mínimo de 42 participações como licitante. Para que uma empresa seja considerada perdedora em uma licitação, ela não pode ter vencido em nenhum item. Todos os valores de parâmetros considerados foram previamente discutidos e definidos com o auxílio de especialistas.

4.4. Trilhas de Auditagem para Alertas de Fraudes

O conceito de trilha de auditagem, definido na Seção 1, pode ser entendido, dentro do contexto da rede social proposta, como uma série de passos (x) para identificar alertas de fraude em licitações. Assim, o resultado de uma trilha pode ser modelado como um conjunto de licitações que descumpriram essas regras. Portanto, o conjunto $T_x = (\hat{G}_1, \hat{G}_2, ..., \hat{G}_j)$, é composto pelas licitações que descumpriram a regra x. Ou seja, a saída da trilha é um conjunto de licitações, portanto é um conjunto de subgrafos da rede social. A Tabela 2 descreve as regras que $\hat{G}_i = (\{\hat{V}_i \cup \hat{\overline{V}}_i\}, \{\hat{E}_i \cup \hat{\overline{E}}_i\})$ precisa atender para que $\hat{G}_i$ seja enquadrada em cada uma das trilhas apresentadas na Tabela 1.

5. Caracterização e Análise da Rede Social Real

Após modelar e construir a rede social entre licitantes e sócios, pode-se utilizar tal estrutura para analisar os relacionamentos entre os indivíduos e verificar possíveis alertas de fraude. Na Seção 5.1, é apresentada uma visão geral da rede a partir de métricas topológicas que a descrevem. Em seguida, a Seção 5.2 analisa a coocorrência de alertas de fraude através da análise de correlação entre as trilhas de auditagem.

5.1. Visão Geral da Rede Social

Nesta seção, são analisados os relacionamentos da rede social para aprofundar o entendimento da dinâmica das conexões entre empresas licitantes e seus sócios. A complexidade

Tabela 2. Regras que $\hat{G}_i = (\{\hat{V}_i \cup \hat{\overline{V}}_i\}, \{\hat{E}_i \cup \hat{\overline{E}}_i\})$ precisa atender para que $\hat{G}_i$ seja enquadrada em cada uma das trilhas de auditagem.

#	Definição Formal da Trilha de Auditagem
	Trilhas de Licitante (Nós)
	$\dot{V}_i \subset \hat{V}_i; \lvert\dot{V}_i\rvert > 0$
T_1	$\dot{V}_i = \{v \mid v[\textit{data\_registro}] < \hat{G}_i[\textit{data\_licitacao}]\}$
T_2	$\dot{V}_i = \{v \mid \lvert v[\textit{sanções}]\rvert > 0\}$
T_3	$\dot{V}_i = \{v \mid v[\textit{status}] \neq \textit{'ATIVO'}\}$
T_4	$\dot{V}_i = \{v \mid v \in \text{perdedores\_frequentes}\}$
T_5	$\dot{V}_i = \{v \mid v \in \text{vencedores\_frequentes}\}$
T_6	$\lvert\dot{V}_i\rvert = 1$
T_7	$\dot{V}_i = \{v \mid \lvert v[\textit{CNAE\_incongruente}]\rvert > 0\}$
	Trilha de Sócio (Nós)
	$\dot{\overline{V}}_i \subset \hat{\overline{V}}_i; \lvert\dot{\overline{V}}_i\rvert > 0$
T_8	$\dot{\overline{V}}_i = \{\overline{v} \mid (\overline{v}[\textit{servidor}] = TRUE) \vee (\lvert v[\textit{parente\_servidor}]\rvert > 0)\}$
	Trilhas de Vínculo (Arestas)
	$\dot{E}_i \subset \hat{E}_i; \lvert\dot{E}_i\rvert > 0$
T_9	$\dot{E}_i = \{(m,n) \in V^2 \mid \forall(m,n) \rightarrow \lvert m[\textit{socios}] \cap n[\textit{socios}]\rvert > 0\}$
T_{10}	$\dot{E}_i = \{(m,n) \in V^2 \mid \forall(m,n) \rightarrow m[\textit{email}] = n[\textit{email}]\}$
T_{11}	$\dot{E}_i = \{(m,n) \in V^2 \mid \forall(m,n) \rightarrow \lvert m[\textit{telefones}] \cap n[\textit{telefones}]\rvert > 0\}$
T_{12}	$\dot{E}_i = \{(m,n) \in V^2 \mid \forall(m,n) \rightarrow m[\textit{endereço}] = n[\textit{endereço}]\}$

Tabela 3. Estatísticas gerais da rede social de licitantes e seus sócios. São apresentados os resultados tanto para a rede completa quanto para seu componente gigante, i.e., o maior componente conectado.

	Completa	C. Gigante		Completa	C. Gigante
Empresas (CNPJ)	54.310	32	Densidade (10^{-4})	0,106	23,690
Pessoas (CPF)	90.347	827	Grau médio	1,531	2,033
Nós (CNPJ e CPF)	144.657	859	CC médio (10^{-1})	0,226	0,008
Arestas	110.750	873	Componentes conectados	43.955	1

CC: Coeficiente de Clusterização

da rede já se revela em seu tamanho: são mais de 144 mil nós, entre empresas e pessoas, que possuem mais de 110 mil arestas entre si. Ilustrar tal rede de forma adequada é uma tarefa desafiadora. Portanto, a Figura 3 apresenta somente a visualização do componente gigante da rede, que compreende o maior conjunto conectado de nós. É importante ressaltar a natureza heterogênea da rede social, que contém nós de diferentes tipos que se conectam em um mesmo contexto.

Em seguida, foram utilizadas métricas topológicas (i.e., obtidas a partir da própria estrutura da rede) para analisar as principais características da rede social: número de nós e arestas, densidade (razão entre a quantidade de arestas existentes e todas as arestas possíveis), grau médio e coeficiente de clusterização (mede a tendência dos nós de se conectarem uns aos outros). Conceitos e definições formais das métricas podem ser acessados em [Barabási 2016].

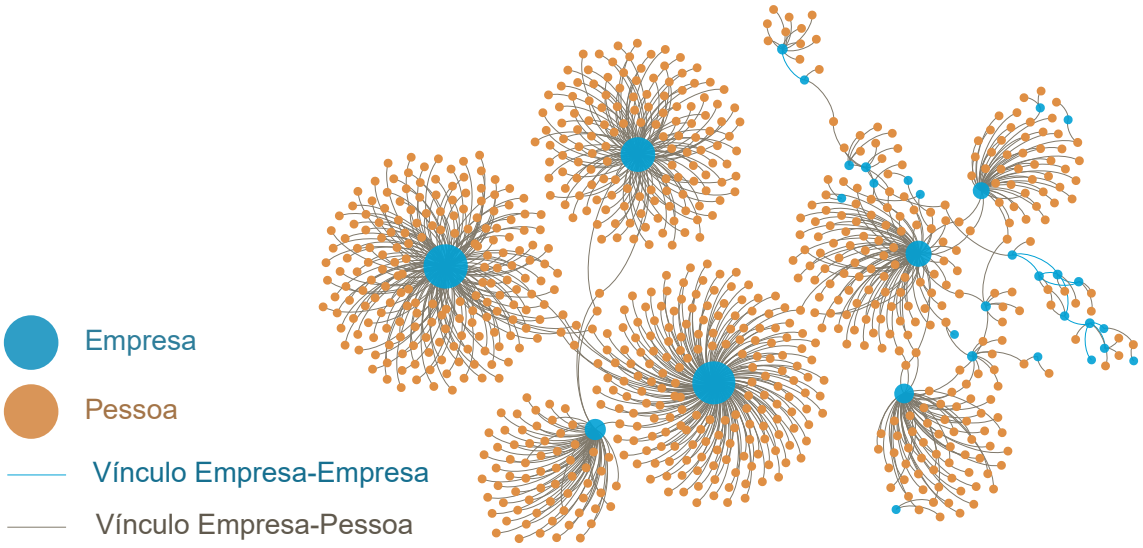

Figura 3. Componente gigante da rede social entre empresas licitantes e seus sócios. Empresas estão conectadas quando possuem vínculos em comum (e.g., endereço, sócios) e participam de um mesmo processo licitatório.

A Tabela 3 apresenta os resultados das métricas topológicas tanto para a rede completa quanto para seu componente gigante. O número de nós e arestas da rede completa revela a complexidade da rede construída. Como esperado, a maior parte dos nós são de pessoas físicas que são sócias de empresas licitantes, uma vez que empresas de médio e grande porte normalmente possuem vários representantes legais. Tal comportamento se repete no componente gigante, que contém 32 empresas e 827 sócios no total.

Ao analisar métricas como o número de arestas, coeficiente de clusterização médio e densidade da rede, percebe-se que a rede social construída é esparsa (valor da densidade próximo a zero), i.e., existem poucas arestas na rede comparadas à quantidade possível. Assim, o fato de relacionamentos na rede serem relativamente raros torna sua análise ainda mais relevante, pois eles podem evidenciar alertas de fraude mais facilmente, além de revelar possíveis esquemas de associações entre licitantes.

Por fim, o número de componentes conectados na rede completa revela uma alta fragmentação entre licitantes. Isso pode significar que as trilhas de auditagem não conseguiram revelar grandes grupos de licitantes que possuam vínculos em comum. Se a rede possuísse poucos componentes conectados, poderia-se imaginar uma grande associação entre empresas licitantes, necessitando uma posterior análise por parte de especialistas. Ainda assim, os relacionamentos dentro dos pequenos grupos existentes na rede devem ser analisados a fim de verificar os alertas de possíveis fraudes. Portanto, análises mais aprofundadas nos resultados das trilhas são necessárias.

5.2. Análise de Correlação entre as Trilhas de Auditagem

Nesta seção, realizamos uma análise de correlação entre as trilhas de auditagem, tanto relacionadas às arestas (pares de licitantes), quanto aos nós (licitantes ou sócios). Para isso, geramos matrizes de correlação para avaliar a ocorrência simultânea entre as trilhas

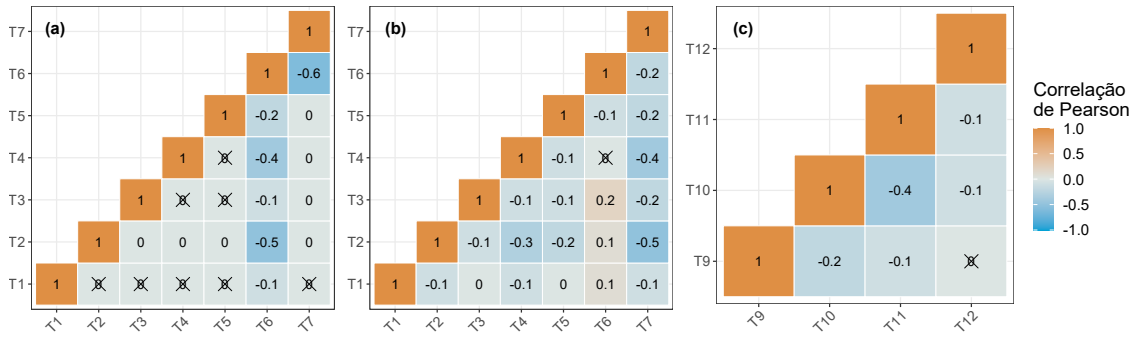

Figura 4. Correlação entre as trilhas relacionadas aos (a-b) licitantes e aos (c) pares de licitante. A matriz (b) é uma versão filtrada da (a), onde foram removidas as ocorrências em que o licitante se enquadra apenas na trilha T_6. O símbolo $\times$ indica uma correlação estatisticamente não significativa (p-valor $\geq$ 0,05).

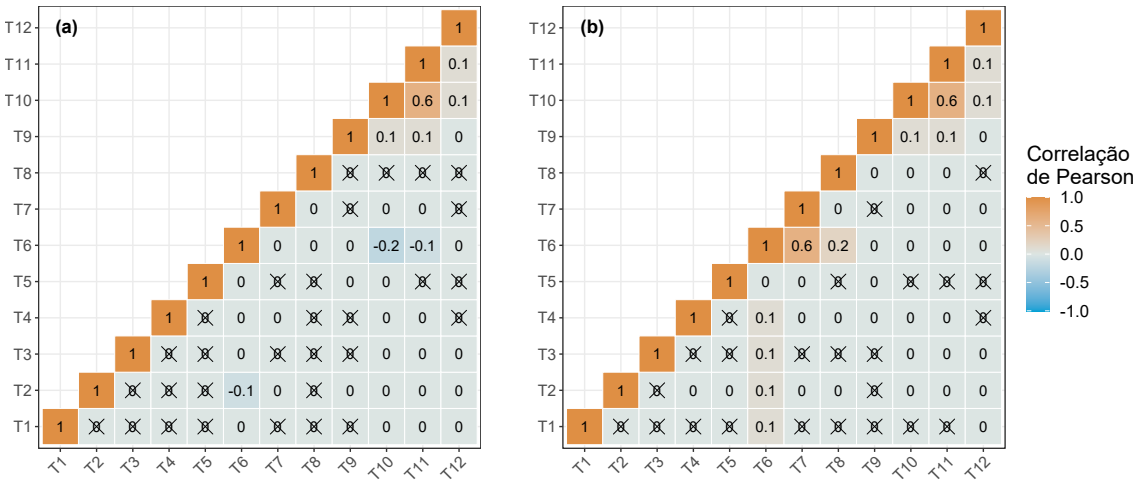

Figura 5. Correlação entre as doze trilhas de auditagem, em nível de licitação. A matriz (b) é uma versão filtrada da (a), onde foram removidas as ocorrências em que a licitação se enquadra apenas na trilha T_6. O símbolo $\times$ indica uma correlação estatisticamente não significativa (p-valor $\geq$ 0,05).

descritas na Tabela 2. Primeiro, analisamos se um *licitante* ou *pares de licitantes* se enquadram simultaneamente em mais de uma trilha. Em seguida, em um nível mais geral, exploramos a coocorrência entre trilhas na mesma licitação. Ao todo, foram avaliadas 212.616 licitações que se enquadram em pelo menos uma das trilhas consideradas.

Correlação em nível de licitante. As Figuras 4(a) e 4(c) mostram as matrizes de correlação de Pearson entre as trilhas referentes aos licitantes e aos pares de licitante, respectivamente. A Figura 4(b) apresenta uma versão filtrada da matriz (a), onde foram removidos os casos que se enquadram apenas na trilha T_6, que investiga licitantes únicos em uma licitação. Essa filtragem foi realizada devido ao volume superior de casos que se enquadram apenas nessa trilha, ocasionando a falta de correlação significativa entre as demais trilhas de licitante (Figura 4(a)). Após a filtragem desses casos, fica evidente o impacto que a trilha T_6 provoca na análise de correlação. Como resultado, podemos notar

que a grande maioria das correlações é estatisticamente significativa e apresentam uma relação negativa fraca a moderada, variando entre -0,1 a -0,5.

Na Figura 4(b), a única correlação forte (-0,5) identificada acontece entre as trilhas T_2 e T_7, que analisam licitações que contenham licitantes com alguma sanção ativa na base do CEIS e licitantes cujo código CNAE é incongruente com a descrição dos itens licitados, respectivamente. Como licitantes cadastrados na base do CEIS não poderiam concorrer em licitações, tal resultado é esperado. De fato, a maioria das correlações negativas, ainda que moderadas, envolve a trilha T_2. O mesmo acontece com a trilha T_7, que apresenta uma relação negativa entre todas as demais trilhas de licitante. Ou seja, no geral, uma empresa que apresenta o código CNAE incongruente não está envolvida em nenhuma outra trilha.

Analisando as trilhas referentes aos pares de licitantes (i.e., T_9 a T_{12}), nota-se novamente que todas as correlações estatisticamente significativas apresentam uma relação negativa fraca a moderada, variando entre -0,1 a -0,4. De fato, a correlação mais evidente (-0,4) acontece entre as trilhas T_{10} e T_{11}, que investigam licitações contendo licitantes com e-mails e telefones em comum, respectivamente. Tal resultado indica que grande parte dos pares de licitantes enquadrados na trilha de e-mails não se enquadram na de telefone, e vice-versa. Ou seja, no geral, empresas licitantes que possuem e-mails em comum, não compartilham o mesmo telefone.

Correlação em nível de licitação. Na Figura 5, foi analisada a correlação entre todas as doze trilhas, agora em nível de licitação. Em outras palavras, foi analisada a ocorrência simultânea das trilhas em uma mesma licitação. Mais uma vez, foi feita a filtragem das licitações que se enquadram apenas na trilha T_6 (Figura 5(b)). No geral, a grande maioria das correlações não é estatisticamente significativa ou não apresentam nenhuma relação, mesmo após a filtragem. A correlação mais evidente é a relação positiva forte (0,6) entre as trilhas T_{10} e T_{11}, indicando que se uma licitação apresentar licitantes distintos com e-mails em comum, há uma grande chance de também apresentar licitantes distintos com número de telefone em comum, e vice-versa.

Após a filtragem, outra correlação positiva (0,6) aparente ocorre entre as trilhas T_6 e T_7. Tal resultado indica que se uma licitação tem licitante único (T_6), ela tem mais chances de ter licitantes cujo código CNAE é incongruente (T_7). Em relação às demais trilhas, a falta de correlação geral revela a inexistência de um padrão entre os alertas, evidenciando o quão desafiadora é a tarefa de análise de possíveis fraudes em licitações.

6. Estudo de Caso de uma Licitação Suspeita de Fraude

Esta seção apresenta um estudo de caso de uma licitação que foi identificada por quatro das trilhas de auditagem descritas neste artigo: T_9, T_{10}, T_{11} e T_{12}. A Figura 6 apresenta diferentes perspectivas de visualização da rede social real dessa licitação de acordo com as trilhas em que ela foi identificada: T_9*:* sócios em comum na Figura 6(a); T_{10}*:* e-mails em comum na Figura 6(b); T_{11}*:* telefones em comum na Figura 6(c); e T_{12}*:* endereços em comum na Figura 6(d). Essa rede é formada por um total de 44 nós, sendo seis nós de empresas (13,64%) e 38 nós de pessoas (86,36%). As arestas representam quatro tipos de vínculos: sócios em comum (sete arestas), e-mail (três arestas), telefone (sete arestas) e endereço (uma aresta). É possível perceber que, como a modelagem da rede é baseada nas

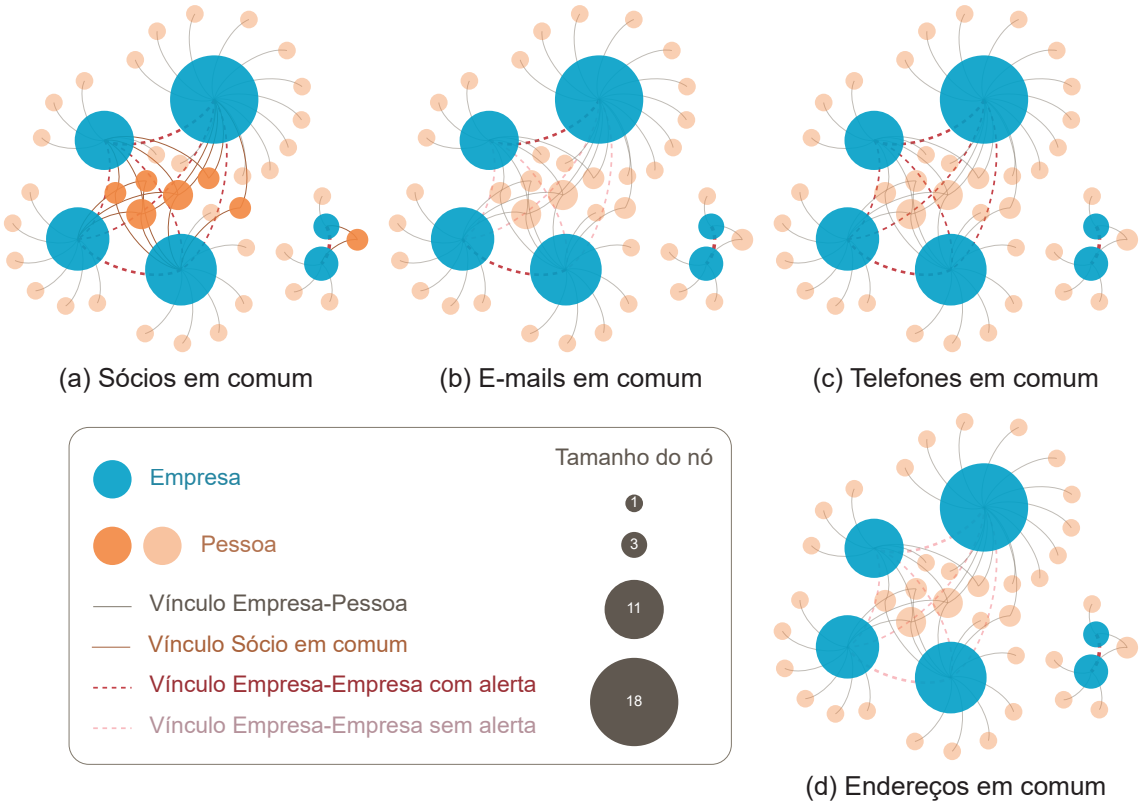

Figura 6. Sub-grafo da rede social real de uma licitação com visualizações diferentes por trilha. O tamanho do nó é baseado no grau do nó.

empresas licitantes, elas tendem a possuir um maior grau na rede em relação às pessoas. Vale destacar que dezoito é o valor do nó de uma empresa com maior grau, e um é o valor do nó de uma empresa e/ou pessoa com menor grau. Neste estudo de caso, empresas com maior grau foram identificadas por mais trilhas de auditagem.

Em relação à trilha T_9 de sócios em comum, observa-se que seis empresas possuem pelo menos um sócio em comum, na cor laranja em destaque na Figura 6(a). Já para a trilha T_{10}, percebe-se três pares de empresas com vínculo Empresa-Empresa com alerta por terem o mesmo e-mail. Para a trilha T_{11}, nota-se que quatro empresas, com maiores graus na rede, possuem telefone em comum entre elas e um par de empresas, com menor grau na rede, também possuem telefone em comum. Por fim, para a trilha T_{12}, apenas um par de empresas, com grau menor na rede, possuem o mesmo endereço. Portanto, essa licitação foi enquadrada nas quatro trilhas devido aos vínculos entre as duas empresas com os dois nós de menor grau. Isso levanta um alerta de fraude na licitação avaliada e, então, recomenda-se um estudo mais aprofundado dos dados referentes a essa licitação.

O estudo de caso revela que a identificação de licitações fraudulentas não é trivial, visto que é necessário a análise de múltiplos aspectos. Apesar da metodologia proposta facilitar na identificação de possíveis fraudes, o papel de especialistas de conteúdo na auditagem de licitações ainda é necessário. Entretanto, frente a tantas licitações, a metodologia permite reduzir os documentos que precisam ser avaliados por especialistas.

7. Ranqueamento de Licitações por Alertas

A combinação dos resultados individuais de cada trilha de auditagem com o valor total pago pelos objetos licitados pode auxiliar no ranqueamento de licitações suspeitas de fraude. Para realizar tal ranqueamento, este artigo apresenta uma abordagem que gera uma pontuação para cada licitação (Seção 7.1), a qual indica o quanto uma licitação pode ser considerada fraudulenta. Em seguida, é definido o grau de risco das trilhas de auditagem (Seção 7.2). Finalmente, é apresentada uma caracterização dos resultados da abordagem de ranqueamento (Seção 7.3).

7.1. Definição da Abordagem de Ranqueamento

Esta seção apresenta a abordagem para ranquear as licitações de acordo com um indicador de risco gerado pelos resultados das trilhas de auditagem em conjunto com o valor total licitado. Esse indicador representa o risco de uma licitação ser objeto de fraude. Especificamente, a abordagem de ranqueamento considera dois aspectos: *(i)* a definição empírica de pesos diferentes para cada trilha de acordo com seu potencial risco de fraude; e *(ii)* o uso do valor total licitado, uma vez que licitações de maior valor são consideradas mais relevantes e têm prioridade na alocação de recursos.

Vale destacar que nos dados de licitações municipais do SICOM (descritos na Seção 3) existem licitações com valores exorbitantes: por exemplo, licitações com valores mais altos que toda a receita do município. Esses casos são prováveis erros de digitação e foram removidos da nossa base de dados para o cálculo do indicador de risco de fraude da licitação.

O indicador de risco de fraude de uma licitação $\hat{G}_i$ pode ser calculado pelo produto escalar do vetor contendo os alertas gerados pelas trilhas para a licitação em questão com o vetor de pesos de cada trilha. Este resultado é depois multiplicado pelo valor normalizado da licitação. Formalmente, tal indicador é dado pela função:

$$I_i\left(Q_{i*}, P, v_i\right) = \left(Q_{i*} \cdot P\right) v_i$$

, onde

- $Q_{202.031\times12} = \{q_{ix} \in \mathbb{N}\}$ é a matriz com a quantidade de alertas de todas as licitações e trilhas, onde as linhas são as licitações e as colunas são as trilhas. A célula dessa matriz, q_{ix}, representa a quantidade de alertas que a licitação i possui na trilha x. Temos um total de 202.031 licitações e 12 trilhas de auditagem consideradas, portanto $i \in [1, 202.031]$ e $x \in [1, 12]$;
- Q_{i*} é uma linha da matriz Q, que representa todos os alertas da licitação i, ou seja, $Q_{i*} = \{q_{i1}, q_{i2}, ..., q_{i12}\}$;
- P é o vetor dos pesos de cada trilha, e o peso atribuído para a trilha x é representado por p_x, ou seja, $P = \{p_x \in \mathbb{R}\}$;
- v_i é o valor total da licitação $\hat{G}_i$ normalizado entre 0 e 1. Note que este valor precisa ser normalizado para que ele não domine numericamente os resultados das trilhas.

A forma para calcular a quantidade de alertas q_{ix} depende do tipo de trilha (apresentados na Tabela 1), são eles: Trilhas de Licitante (Nós), Trilhas de Sócio (Nós) e Trilhas de Vínculo (Arestas). Para as trilhas de licitantes, a quantidade de alertas é o número de nós do tipo empresa que descumprem a regra da trilha. Já para as trilhas de sócio, a quantidade de alertas é o número de nós do tipo pessoa que descumprem a regra da trilha. Finalmente, para as trilhas de vínculo, a quantidade de alertas é o número de arestas que descumprem a regra da trilha. A Equação 1 apresenta a regra para calcular q_{ix} para cada tipo de trilha, baseando-se em suas definições formais (vide Tabela 2).

$$q_{ix} = \begin{cases} |\dot{V}_i|, & \text{se } 1 \leq x \leq 7 \\ |\dot{\overline{V}}_i|, & \text{se } x = 8 \\ |\dot{E}_i|, & \text{se } 9 \leq x \leq 12 \end{cases} \tag{1}$$

7.2. Grau de Risco das Trilhas

Esta seção apresenta uma forma de atribuir os pesos para as trilhas, mas note que esses valores podem ser alterados conforme a necessidade, pois não há um consenso para a atribuição de pesos às trilhas. Por isso, foi atribuído um grau de risco para cada trilha e um valor de peso para cada grau de risco. Esse grau de risco foi atribuído por auditores especializados de acordo com a gravidade do alerta gerado pela trilha de auditagem.

A Tabela 4 apresenta os três graus de risco definidos – *alto*, *médio* e *baixo* –, e descreve a relação deles com cada trilha. A Equação 2 apresenta o peso p_x definido para cada trilha. Em resumo, as trilhas de grau de risco *alto* vão ter peso 1, as trilhas de grau *médio* vão ter peso 0,8 e as trilhas de grau *baixo* vão ter peso 0,6. Esses valores de peso foram definidos empiricamente em conjunto com auditores especialistas do MPMG.

$$p_x = \begin{cases} 1; & \text{se } x \in \{1, 3, 9\} \\ 0,8; & \text{se } x \in \{2, 7, 8, 12\} \\ 0,6; & \text{se } x \in \{4, 5, 6, 10, 11\} \end{cases} \tag{2}$$

7.3. Resultado da Abordagem de Ranqueamento de Licitações

Esta seção apresenta os resultados do uso da abordagem proposta para ranquear as licitações conforme seu indicador de risco (Seção 7.3.1). Também descreve uma análise da influência de cada trilha no resultado do ranqueamento (Seção 7.3.2).

7.3.1. Análise das Licitações no Ranqueamento

Ao todo 202.031 licitações possuem indicador de risco maior que zero e portanto fazem parte do ranqueamento. A Tabela 5 apresenta as top 10 licitações ranqueadas. Nela, a primeira coluna representa a posição da licitação no ranqueamento e as colunas seguintes, T_1 até T_{12}, indicam a quantidade de alertas que a licitação teve em cada uma das 12 trilhas. Finalmente, temos as colunas que representam o preço total de objetos descritos em uma licitação (Preço Licitado) e seu indicador de risco.

Tabela 4. Grau de risco de cada trilha.

Risco	Trilha
Alto	T_1 - Licitante licitando antes de registro T_2 - Licitante licitando com sanção ativa T_3 - Licitante com CNPJ inativo T_9 - Licitantes com sócios em comum
Médio	T_7 - Licitante com CNAE incongruente T_8 - Licitante cujos sócios são ou têm vínculo com servidores públicos T_{12} - Licitantes com endereços em comum
Baixo	T_4 - Licitante perdedor frequente T_5 - Licitante vencedor frequente T_6 - Licitante único T_{10} - Licitantes com e-mails em comum T_{11} - Licitantes com telefones em comum

Tabela 5. Top 10 licitações com maiores valores para o Indicador de Risco.

Posição Licitação	Qtd. Alerta												Preço Licitado	Indicador Risco
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12		
1ª	0	0	1	0	0	0	0	0	0	50	41	13	R$ 3.437.523,60	65,19
2ª	0	0	0	0	0	0	0	0	0	44	38	15	R$ 3.432.317,50	60,36
3ª	0	0	0	0	0	0	0	0	0	47	41	10	R$ 2.710.163,40	47,35
4ª	0	0	0	0	0	0	0	0	0	41	36	8	R$ 2.947.456,20	44,55
5ª	0	0	1	0	0	0	0	0	0	32	24	8	R$ 2.696.717,10	31,77
6ª	0	0	0	0	0	0	0	0	0	39	27	11	R$ 2.139.726,80	29,76
7ª	0	0	1	0	0	0	0	0	0	13	13	2	R$ 3.195.200,00	16,71
8ª	0	0	0	0	0	0	0	0	2	14	18	2	R$ 2.336.762,00	15,31
9ª	0	0	0	0	0	0	0	0	0	16	18	2	R$ 2.352.558,00	14,87
10ª	0	0	0	0	0	0	0	0	0	19	13	0	R$ 2.494.636,88	13,76

A análise do ranqueamento revela que as licitações que estão no topo do ranking possuem elevadas quantidades de alertas para as trilhas T_{10}, T_{11} e T_{12}, apesar dessas duas primeiras trilhas possuírem um peso baixo e a terceira um peso médio (conforme Equação 2), ainda assim essas licitações foram bem ranqueadas por possuírem altas quantidades de alertas nessa trilhas. Por exemplo, para a primeira licitação no Top 10, ela recebeu 50 alertas na trilha e-mails em comum (T_{10}), 41 na trilha de telefones em comum (T_{11}) e 13 na trilha de endereços em comum (T_{12}). Sabemos que isso acontece por um grande número de empresas compartilharem o email e telefone de seus escritórios de contabilidade.

Para avaliar esse ranqueamento, foi realizada sua comparação com uma pequena amostra de 117 licitações já investigadas pelo MPMG e que comprovadamente tiveram algum tipo de fraude. Dessas 117, 18 (15,38%) possuem pelo menos um alerta identificado pela proposta deste trabalho. Portanto, possuem um indicador de risco maior que zero e fazem parte do ranqueamento gerado neste trabalho.

A Tabela 6 apresenta as 18 licitações fraudulentas investigadas por especialistas do MPMG e que possuem um indicador de risco. Essa tabela apresenta a posição que essas licitações ficaram no ranqueamento gerado pela abordagem proposta, mas nenhuma

Tabela 6. Posição no ranqueamento de licitações já investigadas pelo MPMG.

Posição Licitação	Qtd. Alertas T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	Preço Licitado	Indicador Risco
137ª	0	0	0	0	0	0	0	0	0	3	3	0	R$ 3.407.281,40	3,525
461ª	0	0	0	0	0	0	0	0	0	2	2	0	R$ 2.448.252,45	1,688
663ª	0	0	0	0	0	0	0	0	0	2	2	0	R$ 1.855.732,74	1,280
1234ª	0	0	0	0	0	0	0	0	0	2	2	0	R$ 1.194.407,05	0,824
3362ª	0	0	0	0	1	0	0	0	0	2	2	0	R$ 485.500,00	0,419
5909ª	0	0	0	0	0	0	0	0	0	2	2	0	R$ 408.376,05	0,282
13020ª	0	0	0	0	0	1	0	0	0	0	0	0	R$ 839.002,09	0,145
13041ª	0	0	0	0	0	1	0	0	0	0	0	0	R$ 837.664,00	0,144
14726ª	0	0	0	0	0	1	0	0	0	0	0	0	R$ 734.760,00	0,127
16096ª	0	0	0	0	0	0	0	0	0	2	0	0	R$ 333.650,00	0,115
25005ª	0	0	0	0	0	1	0	0	0	0	0	0	R$ 405.546,29	0,070
27546ª	0	0	0	0	0	1	0	0	0	0	0	0	R$ 360.368,24	0,062
31276ª	0	0	0	0	0	0	0	0	0	2	2	0	R$ 77.421,45	0,053
65532ª	0	0	0	0	0	1	0	0	0	0	0	0	R$ 115.000,00	0,020
68227ª	0	0	0	0	0	1	0	0	0	0	0	0	R$ 107.910,00	0,019
73455ª	0	0	0	0	0	1	0	0	0	0	0	0	R$ 96.250,00	0,017
78508ª	0	0	0	0	0	1	0	0	0	0	0	0	R$ 86.311,63	0,015
157824ª	0	0	0	0	0	1	0	0	0	0	0	0	R$ 17.400,00	0,003

delas ficou no Top 10, a licitação de posição mais alta foi 137ª. A justificativa para isso é que existem licitações com quantidade de alertas maiores no topo do ranqueamento, conforme mostra a Tabela 5. Porém, as 4 primeiras licitações da Tabela 6 ficaram acima do percentil 99 no ranqueamento, uma posição relativamente alta.

A amostra de licitações investigadas é muito pequena para termos um resultado muito preciso. Entretanto, há uma intersecção entre as licitações no ranqueamento e as licitações verificadas por especialistas do MPMG. Portanto, podemos observar que o ranqueamento auxilia na priorização das licitações que devem ser investigadas por auditores do MPMG.

7.3.2. Análise do DCG do Ranqueamento

Para medir a influência de cada trilha no resultado do ranqueamento, foi calculado o DCG (*Discounted Cumulative Gain*), que mede o somatório do ganho c_{ix} levando em conta a posição do item em um ranqueamnto [Wang et al. 2013]. A Equação 3 apresenta o cálculo dessa métrica. Para este trabalho, o ganho c_{ix} é a contribuição da trilha x para o Indicador de Risco da licitação i. Conforme mencionado na Seção 7.1, $i \in [1, 202.031]$.

$$DCG_x = \sum_{i=1} \frac{c_{ix}}{log_2(i+1)} \tag{3}$$

$$c_{ix} = q_{ix} \cdot p_x$$

A Figura 7 apresenta o DCG e a quantidade de alertas de cada trilha, ambas

variáveis em escala logarítmica. Também há uma reta para indicar a relação de proporcionalidade entre as duas variáveis. Nessa figura é possível observar que a trilha T_6 (licitantes únicos em uma licitação) é a que mais influencia no ranqueamento. A explicação para isso é que a T_6 é a trilha que mais enquadra licitações, são 202.031 licitações com indicador de risco maior que zero, sendo que 185.254 (91,69%) dessas, se enquadram apenas na trilha T_6. Esse alto volume de licitações enquadradas na T_6 fazem com que essa trilha tenha uma presença muito forte no ranqueamento e, consequentemente, em um alto valor no seu DCG.

O DCG é diretamente proporcional à quantidade de alertas, por isso, é esperado que quanto maior for a quantidade de alertas, maior será o DCG. Portanto, é de se esperar uma linearidade entre os valores do DCG e a quantidade de alertas, conforme mostra a reta na Figura 7. As trilhas de nós (T_1 até T_8) seguem esse padrão, mas as trilhas de vínculo (T_9 até T_{12}), possuem um DCG, proporcionalmente, maior que suas quantidades de alertas.

As trilhas de vínculo representam empresas, participantes da mesma licitação, que possuem algum vinculo em comum. Os vínculos que avaliamos são socio, e-mail, telefone e endereço em comum. Se duas empresas estiverem situadas no mesmo escritório, provavelmente terão as mesmas informações cadastrais. Isso é evidenciado pelo fato das trilhas T_{10} e T_{11} possuírem uma correlação moderada-forte (0,6) entre si (apresentado na Figura 5(b)), ou seja, na maioria dos casos, as licitações enquadradas na trilha T_{10} também são enquadradas na trilha T_{11}, essa combinação eleva o indicador de risco dessas licitações aumentando suas posições no ranqueamento e melhorando os valores de DCG das trilhas T_{10} e T_{11}.

Já as trilhas T_{09} e T_{12}, possuem uma correlação fraca, porem positiva com as trilhas T_{10} e T_{11}, conforme consta na Figura 5(b). Isso faz com que algumas licitações enquadradas nas trilhas T_{09} e T_{12} também sejam enquadradas em T_{10} ou T_{11}, além do fato de T_{09} e T_{12} possuírem graus de risco alto e médio, respectivamente, fazendo com o peso dessas trilhas seja maior. Esses fatores aumentam o indicador de risco das licitações enquadradas nessa trilhas e suas posições no ranqueamento, consequentemente melhorando o DCG de T_{09} e T_{12}.

8. Limitações e Desafios

Para as trilhas $T_3, T_9, T_{10}, T_{11}, T_{12}$ existe a limitação da falta de dados históricos, pois não há informações de sócios, telefones, endereços, etc., no momento em que a licitação aconteceu. Há apenas os valores que constavam na base do Serpro em agosto de 2021. Essa falta de dados pode resultar em falsos positivos nos resultados das trilhas. Algumas empresas que usam serviços de escritório de contabilidade cadastram, junto a receita federal, o e-mail, telefone e endereço do escritório de contabilidade ao invés dos dados da empresa. Isso gera ruído nos resultados das trilhas T_{10}, T_{11}, T_{12}.

Além disso, a falta de informações sobre os licitantes perdedores das licitações estaduais pode impactar nos resultados das trilhas sobre elas. Ou seja, as trilhas vão considerar somente os vencedores como participantes. No entanto, esta limitação não invalida a aplicação das trilhas nessas licitações, uma vez que ainda é possível identificar

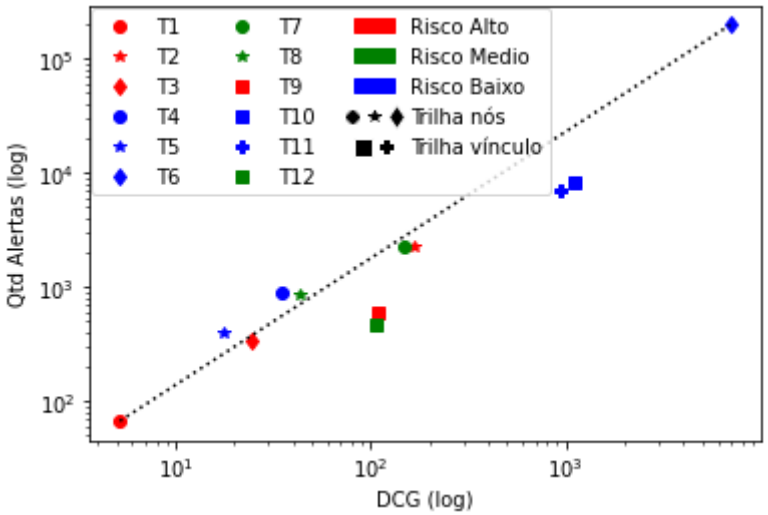

Figura 7. DCG (log) x Quantidade de Alertas por trilha (log) de cada uma das trilhas. Há também uma reta de referencia para observarmos a proporcionalidade entre as duas variáveis.

indícios de irregularidades relacionadas aos licitantes vencedores.

Em relação à abordagem de ranqueamento, uma limitação é a pequena quantidade de dados rotulados para realizar uma avaliação mais precisa do resultado. Outra limitação é que os pesos para as trilhas foi definido por um especialista, o ideal é que mais especialistas participassem da atribuição destes pesos. Além disso, os valores dos pesos são atribuídos para aumentar ou diminuir a relevância de cada trilha no cálculo do indicador de risco e não há atribuição otimizada para esses pesos, visto que cada especialistas tem interesse em investigar determinados aspectos do processo licitatório e, consequentemente, a relevância de cada trilha será considerada diferente. Além disso, o grau de importância do resultado de uma trilha pode variar de acordo com o momento, por exemplo: é recebida uma denúncia de empresas participando de licitações junto com outras empresas de mesmo sócio. No momento de investigação dessa denúncia, a trilha "T09 - Licitações contendo licitantes com representantes sócios em comum" terá uma importância muito maior.

9. Conclusões e Trabalhos Futuros

Este artigo apresentou um conjunto de trilhas de auditagem para identificação de licitações públicas que são suspeitas de fraude. No total, foram descritas 12 trilhas, das quais sete são referentes ao licitante, uma avalia os sócios e quatro são relacionadas ao tipo de vínculo. Essas trilhas de auditagem foram então modeladas como uma rede social, na qual os nós são empresas licitantes ou os sócios dessas empresas, e as arestas representam vínculos entre as empresas e/ou sócios. Essa modelagem possibilitou identificar características de vínculos entre empresas licitantes e seus sócios que podem indicar licitações suspeitas de fraude.

Este trabalho também propôs uma abordagem que ranqueia as licitações com intuito de indicar as que possuem maior risco de serem fraudulentas. Esse ranqueamento é

baseado nos resultados gerados pelas doze trilhas de auditagem. Em geral, os resultados mostram que nossas propostas são promissoras, principalmente, em relação a três aspectos, são eles: *(i)* revela ser possível utilizar a metodologia para filtrar os dados; *(ii)* reduz o volume de dados a serem analisados por especialistas; e *(iii)* indica uma possível ordem de prioridade para análise das licitações. Além disso, essas descobertas geram subsídio para elaboração de algoritmos capazes de classificar uma licitação como fraudulenta ou não, de forma a auxiliar no combate à corrupção.

Como trabalhos futuros, planejamos melhorar a abordagem de ranqueamento proposta de forma a aumentar a granularidade da escala que define o grau de risco de uma trilha. Também pretendemos utilizar as características identificadas nas trilhas de auditagem e na abordagem de ranqueamento como entrada para um algoritmo de classificação para licitações públicas.

Agradecimentos. Ao Ministério Público de Minas Gerais (MPMG) pelo apoio através do Projeto Capacidades Analíticas. Ao CNPq, CAPES e Fapemig pelo apoio aos pesquisadores envolvidos.

Referências

[Abidi et al. 2021] Abidi, W. U. H. et al. (2021). Real-time shill bidding fraud detection empowered with fussed machine learning. *IEEE Access*, 9:113612–113621.

[Andrade et al. 2020] Andrade, P. H. M. A. et al. (2020). Auditing government purchases with a multicriteria anomaly detection strategy. *J. Inf. Data Manag.*, 11(1).

[Anowar and Sadaoui 2019] Anowar, F. and Sadaoui, S. (2019). Multi-class ensemble learning of imbalanced bidding fraud data. In *Canadian AI*, volume 11489 of *Lecture Notes in Computer Science*, pages 352–358. Springer.

[Aquino Jr et al. 2019] Aquino Jr, G. S. d., Jacob, E., Henrique, G., Guerethes, J., and de Oliveira Silva, R. (2019). Dados abertos para o fomento da transparência e inovação: o caso da ufrn. *iSys-Brazilian Journal of Information Systems*, 12(2):39–59.

[Araújo et al. 2021] Araújo, J. L. et al. (2021). Caracterização de "Caminhos mais prováveis" em uma rede complexa de processos jurídicos. In *BraSNAM*, pages 44–54, Porto Alegre, Brasil. SBC.

[Barabási 2016] Barabási, A.-L. (2016). *Network science*. Cambridge University Press.

[Brum et al. 2022] Brum, P., Cândido Teixeira, M., Vimieiro, R., Araújo, E., Meira Jr, W., and Lobo Pappa, G. (2022). Political polarization on twitter during the covid-19 pandemic: a case study in brazil. *Social Network Analysis and Mining*, 12(1):1–17.

[Costa et al. 2022] Costa, L., Reis, A., Bacha, C., Oliveira, G., Silva, M., Teixeira, M., Brandão, M., Lacerda, A., and Pappa, G. (2022). Alertas de fraude em licitações: Uma abordagem baseada em redes sociais. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 37–48, Porto Alegre, RS, Brasil. SBC.

[Elshaar and Sadaoui 2020] Elshaar, S. and Sadaoui, S. (2020). Detecting bidding fraud using a few labeled data. In *ICAART*, pages 17–25. SCITEPRESS.

[Florentino et al. 2022] Florentino, É. S., Goldschmidt, R. R., and Cavalcanti, M. C. (2022). Identificando suspeitos de crimes por meio de interações implícitas no youtube. *iSys-Brazilian Journal of Information Systems*, 15(1):3–1.

[Ganguly and Sadaoui 2018] Ganguly, S. and Sadaoui, S. (2018). Online detection of shill bidding fraud based on machine learning techniques. In *IEA/AIE*, volume 10868 of *Lecture Notes in Computer Science*, pages 303–314. Springer.

[Grace et al. 2016] Grace, E. et al. (2016). Detecting fraud, corruption, and collusion in international development contracts: The design of a proof-of-concept automated system. In *IEEE BigData*, pages 1444–1453. IEEE Computer Society.

[Kansaon et al. 2019] Kansaon, D. P., Brandao, M. A., and de Paula Pinto, S. A. (2019). Análise de algoritmos de classificação para detecção de emoções em tweets em português brasileiro. *iSys-Brazilian Journal of Information Systems*, 12(3):116–138.

[Lima et al. 2020] Lima, M. et al. (2020). Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach. In *EMNLP*, pages 1580–1588, Online. ACL.

[Oliveira et al. 2022a] Oliveira, G. P., Reis, A. P., Freitas, F. A., Costa, L. L., Silva, M. O., Brum, P. P., Oliveira, S. E., Brandão, M. A., Lacerda, A., and Pappa, G. L. (2022a). Detecting inconsistencies in public bids: An automated and data-based approach. In *Brazilian Symposium on Multimedia and Web*, pages 182–190.

[Oliveira et al. 2022b] Oliveira, G. P., Reis, A. P., Mendes, B. M., Bacha, C. A., Costa, L. L., Canguçu, G. L., Silva, M. O., Caetano, V., Brandão, M. A., Lacerda, A., et al. (2022b). Ferramentas open-source de qualidade de dados para licitações públicas: Uma análise comparativa. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 116–127. SBC.

[Óskarsdóttir et al. 2022] Óskarsdóttir, M., Ahmed, W., Antonio, K., Baesens, B., Dendievel, R., Donas, T., and Reynkens, T. (2022). Social network analytics for supervised fraud detection in insurance. *Risk Analysis*, 42(8):1872–1890.

[Park and Kim 2020] Park, C. H. and Kim, K. (2020). E-government as an anti-corruption tool: Panel data analysis across countries. *International Review of Administrative Sciences*, 86(4):691–707.

[Pereira and Murai 2021] Pereira, R. and Murai, F. (2021). Quão efetivas são redes neurais baseadas em grafos na detecção de fraude para dados em rede? In *BraSNAM*, pages 205–210, Porto Alegre, Brasil. SBC.

[Ralha and Silva 2012] Ralha, C. G. and Silva, C. V. S. (2012). A multi-agent data mining system for cartel detection in brazilian government procurement. *Exp. Syst. Appl.*, 39(14):11642–11656.

[Santoro et al. 2020] Santoro, F. M., Revoredo, K. C., Costa, R. M., and Barboza, T. M. (2020). Process mining techniques in internal auditing: A stepwise case study. *iSys-Brazilian Journal of Information Systems*, 13(4):48–76.

[Silva et al. 2022] Silva, M. O., Paula, A. F., Oliveira, G. P., Vaz, I. A., Hott, H., Gomide, L. D., Reis, A. P., Mendes, B. M., Bacha, C. A., Costa, L. L., et al. (2022). Lipset: Um

conjunto de dados com documentos rotulados de licitações públicas. In *Anais do IV Dataset Showcase Workshop*, pages 13–24. SBC.

[Velasco et al. 2021] Velasco, R. B. et al. (2021). A decision support system for fraud detection in public procurement. *Int. Trans. Oper. Res.*, 28(1):27–47.

[Vlasselaer et al. 2015] Vlasselaer, V. V. et al. (2015). AFRAID: fraud detection via active inference in time-evolving social networks. In *ASONAM*, pages 659–666. ACM.

[Vlasselaer et al. 2017] Vlasselaer, V. V. et al. (2017). Gotcha! network-based fraud detection for social security fraud. *Manag. Sci.*, 63(9):3090–3110.

[Wang et al. 2013] Wang, Y., Wang, L., Li, Y., He, D., Chen, W., and Liu, T.-Y. (2013). A theoretical analysis of ndcg ranking measures. In *Proceedings of the 26th annual conference on learning theory (COLT 2013)*, volume 8, page 6.