PEP Scan - A Tool for Assisting in the Identification of Potential Improper Benefits Receipt by PEPs in Brazil

Michel de O. Guijarro[1], **Claudio F. M. Toledo**[1], **Valter V. Camargo** [2]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Caixa Postal 668 – 13.560-970 – São Carlos – SP – Brazil

[2]Departamento de Computação – Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13.565-905 – São Carlos – SP – Brazil

michelguijarro@alumni.usp.br, claudio@icmc.usp.br
valtervcamargo@ufscar.br

Abstract.
Research Context: *PEP Scan focuses on the use of technology to combat corruption in Brazil. Its primary objective is the monitoring of Politically Exposed Persons (PEP), aiming to prevent the misuse of public funds and strengthen transparency.* ***Scientific and/or Practical Problem:*** *There is a critical gap in public tools for integrating and analyzing heterogeneous government data. The core challenge lies in adapting data mining techniques to identify fraud, with a specific focus on uncovering complex connections.* ***Proposed Solution and/or Analysis:*** *The proposed solution is PEP Scan, a web-based tool that integrates public data. It employs a Breadth-First Search (BFS) algorithm and Visual Data Mining (VDM) techniques to map suspicious connections, enabling the exploration of patterns through an interactive graph interface.* ***Related IS Theory:*** *Pep Scan aligns with socio-technical IS theory by integrating technical innovations and user-centric needs.* ***Research Method:*** *The methodology involved the development of a prototype that integrated and processed public databases. The research core was the adaptation of the BFS algorithm to map connections. Two kinds of evaluation were performed: an analysis of the architectural quality and another related to usability, employing the heuristics proposed by Jakob Nielsen.* ***Summary of Results:*** *PEP Scan demonstrated its effectiveness by identifying PEPs who received benefits unduly, including over 400 city councilors. The tool also revealed indirect and complex connections, validating its capability to uncover non-trivial patterns.* ***Contributions and Impact to IS area:*** *This research provides a replicable model and a practical tool for combating corruption. The work contributes to the Information Systems field by demonstrating how technical rigor and a focus on usability can promote citizen science and strengthen public transparency.*

1. Introdução

O sistema apresentado neste artigo tem como objetivo endereçar a crescente preocupação com esquemas de corrupção, fraude e desvio de recursos públicos no Brasil, frequentemente envolvendo atores políticos [Kerche and Marona 2022]. A corrupção constitui um desafio global que compromete a efetividade das políticas públicas, distorce a alocação de

recursos e corrói a confiança da sociedade nas instituições democráticas. No caso brasileiro, o problema assume proporções graves: segundo o Índice de Percepção da Corrupção (IPC) de 2023, o país ocupa a 104ª posição entre 180 nações, com pontuação de 36 em 100, abaixo da média global de 43 pontos [Transparência Internacional 2024]. Ainda mais relevante para este estudo é o fato de que, entre os principais vetores de risco figuram as Pessoas Politicamente Expostas (PEP), definidas pela Resolução nº 29/2017 da Controladoria-Geral da União (CGU) como indivíduos que ocupam ou ocuparam, nos últimos cinco anos, cargos públicos de alto escalão, tais como ministros, parlamentares, diretores de estatais e altos funcionários do Judiciário [Brasil 2017]. Apesar do arcabouço normativo, a ausência de ferramentas públicas e automatizadas para monitorar transações e relações desses agentes limita a eficácia das políticas de prevenção e combate à corrupção.

Para enfrentar esse desafio, técnicas de mineração de dados e teoria dos grafos são promissoras na identificação de padrões ocultos associados a fraude e lavagem de dinheiro. Abordagens baseadas em grafos permitem revelar estruturas relacionais e comportamentos anômalos que frequentemente escapam a métodos estatísticos tradicionais [Aggarwal et al. 2015]. Tais metodologias já foram aplicadas com êxito em domínios como finanças, saúde e transporte, evidenciando sua utilidade para a detecção de esquemas complexos. No contexto de PEPs, o mapeamento de redes de relacionamento com familiares, associados e empresas é reconhecido por organismos internacionais, como o GAFI/FATF, como componente essencial de políticas de *compliance* e de monitoramento bancário, além de apoiar investigações jornalísticas voltadas a coibir crimes financeiros e preservar a confiança pública [Sharman 2009, Financial Action Task Force (FATF) 2013, King and Walker 2016]. Apesar da reconhecida eficácia dessas técnicas e da relevância do problema, ainda permanece a lacuna de uma base pública, centralizada e padronizada para o monitoramento de PEPs. Isso ocorre porque as informações encontram-se dispersas em múltiplas fontes abertas e governamentais, como os portais de transparência, diários oficiais, Tribunal Superior Eleitoral (TSE) e Congresso Nacional, exigindo integração e curadoria sistemáticas.

Essa é a lacuna que este trabalho se propõe a preencher ao introduzir PEP Scan, um protótipo de aplicação web que combina mineração de dados, análise de grafos e visualização interativa para identificar potenciais recebimentos indevidos de benefícios e conexões suspeitas envolvendo PEPs no contexto brasileiro. Em termos metodológicos, mineração de dados é entendida como o processo de extração de padrões úteis em grandes volumes de dados, incluindo regras de associação, classificação, agrupamento e detecção de anomalias [Han et al. 2022]. A literatura destaca sua relevância frente ao crescimento do volume e da complexidade informacional [Aggarwal et al. 2015, Witten et al. 2005]. Complementarmente, a Teoria de Grafos fornece a estrutura formal para representar e analisar relacionamentos como redes de vértices e arestas, permitindo caracterizar dependências, modularidade e centralidades em sistemas complexos [West 2001, Borgatti et al. 2009, Newman 2018]. A integração dessas duas frentes possibilita construir modelos relacionais escaláveis e interpretáveis para o monitoramento de PEPs com base em dados públicos heterogêneos.

As principais contribuições deste estudo são de de natureza conceitual, metodológica e tecnológica:

- Consolidação e *pipeline* de integração de múltiplas bases públicas relevantes (registros de PEPs, beneficiários do Auxílio Emergencial e Bolsa Família, e dados do CNPJ), com limpeza, padronização e enriquecimento de atributos voltados à análise relacional em larga escala;
- Modelagem de dados em grafos, com definição de esquemas de nós e arestas que capturam relações pessoa–pessoa, pessoa–empresa e pessoa–benefício, além de algoritmos para priorização de entidades e identificação de padrões suspeitos;
- Desenho e implementação do PEP Scan, um protótipo web que integra mineração de dados e algoritmos em grafos com visualizações interativas para exploração e auditoria de redes associadas a PEPs; e
- Uma estratégia de avaliação do protótipo, contemplando critérios de cobertura (das fontes integradas) e utilidade analítica (capacidade de priorização e explicabilidade dos indícios).

O objetivo central foi o desenvolvimento do PEP Scan; uma ferramenta pública e automatizada para descoberta e análise de relações envolvendo PEPs, capaz de apoiar prevenção e detecção de irregularidades. Para alcançar isso, o trabalho abordage três itens principais: i) trata e integra bases públicas; ii) define atributos e procedimentos de mineração que viabilizem a análise de conexões em escala e iii) entrega um protótipo funcional passível de avaliação sistemática.

Ao articular mineração de dados e teoria dos grafos em um fluxo unificado, o estudo busca contribuir para o fortalecimento da governança, da transparência e da integridade nas instituições públicas brasileiras.

2. Fundamentação Teórica

Os fundamentos que sustentam o PEP Scan foram articulados em três pilares principais: (i) processo da Descoberta de Conhecimento em Bases de Dados (DCBD) e técnicas de mineração de dados aplicadas a dados públicos heterogêneos; (ii) a análise de grafos voltada à detecção de anomalias e ao Prevenção à Lavagem de Dinheiro (PLD); e (iii) princípios de visualização analítica e escolhas tecnológicas para visualização interativa na web. Adicionalmente, são discutidos trabalhos relacionados clássicos e recentes e, ao final, é sintetizado como o presente trabalho se diferencia dos demais.

2.1. DCBD e Mineração de Dados

O trabalho seminal de Fayyad formaliza a DCBD como um processo iterativo que abrange seleção, pré-processamento, transformação, mineração e interpretação/avaliação de padrões [Fayyad 1997]. Nesse enquadramento, a mineração de dados corresponde ao uso de algoritmos para extrair padrões úteis, sendo parte de um ciclo mais amplo de preparação e validação.

Em bases públicas brasileiras como registros de PEPs, benefícios sociais e CNPJ, sobressaem os desafios de heterogeneidade informacional, a semântica dos dados abertos e a transformação para análise relacional, estendendo engenharia de características para representações em grafo [Han et al. 2022].

Nesse sentido, além das tarefas clássicas (classificação, agrupamento e detecção de anomalias), janelas temporais móveis são empregadas para modelar dependências entre eventos, o que permite capturar relações do tipo apontado na Equação (1), onde

α é definido de acordo com o tempo em que uma pessoa é considerada PEP, mesmo após o seu mandato, ampliando abordagens baseadas apenas em coocorrências estáticas [Han et al. 2022]. Além disso, a interpretação de padrões em contextos complexos é apoiada por princípios da Mineração Visual de Dados (MVD) [Keim and Kriegel 2002], integrados ao fluxo DCBD para sustentar exploração iterativa e auditável.

$$\text{PEP}(X) \wedge \text{Benefício do Governo}(Y) \xrightarrow{\Delta t \leq \alpha \text{ meses}} \text{Possível Recebimento Ilegal} \quad (1)$$

2.2. Análise de Grafos

A aplicação de técnicas de análise de grafos à identificação de atividades fraudulentas e ao combate à lavagem de dinheiro constitui um campo consolidado na literatura, mas com desafios específicos quando transposto para o contexto de políticas públicas.

Neste contexto, a sistematização de abordagens de Detecção de Anomalias Baseadas em Grafos (DABG) destaca sua eficácia superior a métodos estatísticos tradicionais em cenários com relações complexas entre entidades [Pourhabibi et al. 2020]. Enquanto o estudo foca em transações financeiras, o PEP Scan adapta esses princípios para redes de influência política, onde os nós representam atores (PEPs, beneficiários, empresas e benefícios) e arestas codificam relações de vínculos, como de socios de empresas e beneficiários de programas do governo.

A inovação do PEP scan reside na operacionalização do DABG para dados não financeiros, seguindo a premissa de que grafos capturam a natureza intrínseca de esquemas de fraude [Pourhabibi et al. 2020], mas estendendo-a a relações de poder político. Adicionalmente, como a detecção eficaz de fraudes requer a combinação de algoritmos de exploração de grafos com métricas topológicas [Akoglu et al. 2015], o PEP Scan implementa essa abordagem através da Busca em Largura adaptada. Ao contrário de métodos baseados em aprendizado profundo que sacrificam interpretabilidade, a Busca em Largura foi escolhida por permitir a identificação de caminhos entre PEPs e beneficiários de acordo com a demanda do usuário (processo iterativo), balanceando profundidade e custo computacional. Outro motivo foi a capacidade da ferramenta gerar explicações auditáveis para relações sinalizadas, crítico em processos de controle público. Portanto, através do PEP Scan, pode-se realizar o *download* dos grafos de conexões suspeitas para futuras investigações mais detalhadas por órgãos de controle.

2.3. Visualização e Ferramentas Web para Grafos

A efetividade de sistemas analíticos depende de traduzir estruturas e medidas em representações visuais interativas que preservem contexto e escalem com o volume de dados. O Gephi popularizou fluxos de trabalho de análise exploratória e métricas com *layouts* interativos, embora seu foco em ambiente local imponha barreiras de acessibilidade e automação [Bastian et al. 2009].

Em contraste, o PEP Scan adota bibliotecas nativas da web para grafos interativos, como integração Python $\rightarrow$ HTML/JS), favorecendo os seguintes aspectos: (i) publicação e compartilhamento por meio de artefatos portáveis, (ii) encadeamento direto com *pipelines* de dados (sem exportações manuais intermediárias) e (iii) filtragem alinhadas a tarefas de auditoria. Essa estratégia emprega, no contexto de políticas públicas, a MVD com

ênfase em transparência, reprodutibilidade e colaboração entre atores técnicos e órgãos de controle [Keim and Kriegel 2002].

Em suma, o PEP Scan apoia-se em um fluxo DCBD adaptado a dados públicos heterogêneos com janelas temporais móveis. Para isso, emprega modelagem e exploração em grafos centradas em explicabilidade e priorização auditável, ancoradas em pesquisas de detecção de anomalias e fraude [Akoglu et al. 2015, Pourhabibi et al. 2020]. O sistema também utiliza recursos de vizualização interativa web para democratizar o acesso a evidências relacionais. Diferente de abordagens focadas em transações financeiras proprietárias, o PEP Scan integra múltiplas fontes públicas brasileiras, enfatiza rastreabilidade desde o projeto (caminhos e subgrafos exportáveis) e prioriza transparência e reprodutibilidade, requisitos críticos para governança e controle social em políticas públicas.

3. PEP Scan: Materiais e Métodos

Esta seção descreve as bases de dados, tecnologias e algoritmos empregados no desenvolvimento do PEP Scan, além da metodologia aplicada para tratamento e acesso eficiente aos dados.

3.1. Bases de Dados Públicas Utilizadas

Para o desenvolvimento do projeto, as bases de dados usadas no desenvolvimento considerou a disponibilidade pública e relevância ao problema. A Tabela 1 mostra as quatro bases utilizada, todas elas obtidas em arquivos CSV, e algumas informações básicas de cada uma. Nessa linha, a Tabela 2 apresenta, como exemplo, a estrutura da base PEP[1] por meio de um recorte de seus dados [GCU 2023]. Este recorte demonstra que a base possui granularidade necessária para análises temporais e de sobreposição de benefícios.

Tabela 1. Bases de Dados Utilizadas na Ferramenta

Nome	Tipo	Período	Linhas	Tamanho
PEP	*full*	mar./2023	133.829	17,3 MB
Bolsa Família	incremental	jan./2013 – nov./2021	499.201.385	54 GB
Aux. Emergencial	incremental	abr./2022 – fev./2023	736.886.267	120,3 GB
CNPJ	*full*	mar./2023	35.050.791	3,4 GB

Tabela 2. Amostra - Dados de PEP

CPF	Nome_PEP	Função	Órgão	Dt_Iní_Exerc	Dt_Fim_Exerc
***.531.324-**	MARCELO	CAS-1	21/12/2019	Não informada	Não informada
***.708.767-**	AARAO	PREFEI	01/01/2017	31/12/2020	31/12/2025
***.810.311-**	AGUIAR	VEREAD	01/01/2021	31/12/2024	31/12/2029

Nota - o nome do PEP foi limitado nesta amostra apenas para fins de apresentação desta tabela.

Adicionalmente, para identificar potenciais irregularidades, é apresentado na Tabela 3 amostras da base de dados do programa Bolsa Família[2]. Paralelamente, a Tabela 4

[1] https://portaldatransparencia.gov.br/download-de-dados/pep

[2] https://portaldatransparencia.gov.br/download-de-dados/bolsa-familia-pagamentos

apresenta alguns registros da base de Auxílio Emergencial[3]. Tais bases apresentam atributos operacionais (mês, UF/município), identificadores do favorecido e do responsável, enquadramento, parcela, valor e observações.

Tabela 3. Amostra - Auxílio Emergencial (Pagamentos)

Mês	UF	Cód_IBGE	Município	CPF	Beneficiário
202101	TO	1722107	Xambioá	***.752.901-**	Eislom
202101	TO	1722107	Xambioá	***.916.431-**	Josué
202101	TO	1722107	Xambioá	***.026.071-**	Leomar

Nota - o nome do beneficiário foi limitado nesta amostra apenas para fins de apresentação desta tabela.

Tabela 4. Amostra - Dados de Bolsa Família (Pagamentos)

Mês	UF	Cód_SIAFI	Município	CPF	Favorecido
202111	AC	645	Bujari	***.842.612-**	Maria
202111	AL	2713	Batalha	***.806.204-**	Thiago
202111	AL	2791	Maravilha	***.664.784-**	Cremilda

Nota - o nome do favorecido foi limitado nesta amostra apenas para fins de apresentação desta tabela.

A base de CNPJ[4] (Empresas, Estabelecimentos, Classificação Nacional de Atividades Econômicas, Qualificações e Sócios) sustenta a identificação de vínculos societários entre pessoas e empresas. Desse modo, a Tabela 5 exemplifica variáveis (identificação do sócio, qualificação, datas, representante legal) críticas ao mapeamento de relações PF–PJ.

Tabela 5. Amostra - Dados de CNPJ (Sócios)

Nome Sócio	CNPJCPF	Qualific.	Dt Iní. Soc.	CPF Rep. Legal	Rep. Legal
ALDIR	***038019**	22	20221215	***000000**	NaN
LUCAS	***474319**	49	20221215	***000000**	NaN
FUJIYAMA	45934113000100	22	20221215	***971367**	COELHO

Nota: os nomes do sócio e do representante legal foram limitados nesta amostra para fins de apresentação.

3.2. Tecnologias e Arquitetura de Dados

A pilha de tecnologias escolhidas para o desenvolvimento privilegiou ferramentas abertas, escaláveis e de ampla adoção: Python para coleta e pré-processamento; Pandas/polars para transformação; SQLite3 como armazenamento relacional leve e Colab/Drive para contornar restrições de armazenamento e processamento locais. Nesse sentido, a Tabela 6 apresenta a principais bibliotecas utilizadas nesse projeto, e suas versões. O ambiente de desenvolvimento escolhidos foi o VS Code e utilizando-se versionamento Git/GitHub para assegur rastreabilidade e colaboração.

No nível de aplicação web, o *back-end* foi implementado com Flask e SQLAlchemy/Flask-SQLAlchemy (camada de persistência), e o módulo de redes empregou `NetworkX` (cálculo/estrutura) e `pyvis` (renderização HTML/JS). Em produção, `gunicorn`/`uvicorn` foram considerados como servidores de aplicação.

[3] `https://portaldatransparencia.gov.br/download-de-dados/auxilio-emergencial`

[4] `https://arquivos.receitafederal.gov.br/index.php/s/YggdBLfdninEJX9`

Tabela 6. Especificações de Software e Versões Utilizadas

Categoria	Bibliotecas e Versões Principais
Backend & Servidor Web	Flask (2.2.5), Flask-SQLAlchemy (3.1.1), Gunicorn (20.1.0), Uvicorn (0.15.0), Werkzeug (2.2.2), Jinja2 (3.1.4)
Processamento de Dados	Pandas (2.2.2), Numpy (2.1.1), Scipy (1.14.1)
Grafos & Redes	NetworkX (3.3), Pyvis (0.3.2)
Visualização & Geoespacial	Matplotlib (3.9.2), Seaborn (0.13.0), Geopandas (1.0.1), Shapely (2.0.6)
Banco de Dados & ORM	SQLAlchemy (2.0.30)
Ambiente & Utilitários	Ipython (8.31.0), Matplotlib-inline (0.1.7)

Nota - As versões listadas referem-se ao ambiente de execução utilizado nos experimentos para garantir a reprodutibilidade.

No sentido de averiguar a qualidade das bases de dados públicas utilizadas, realizou-se uma análise exploratória. Essa análise revelou dados com problemas, como valores ausentes, CPFs/CNPJs parciais, nomes truncados e heterogeneidade de formatos. Diante disso, foram realizados procedimentos de limpeza de dados, desconsiderando registros com dados nulos ou vazios, desconsideração de possíveis dados duplicados, aplicando expressões regulares e normalização semântica (cargos, órgãos e atributos) para minimizar as ambiguidades.

Para a criação dos nós dos grafos, adotou-se a concatenação **CPF/CNPJ parcial + Nome Completo** como identificador operacional. Entretanto, por mais que essa contatenação é considerada uma chave forte, não há garantia de ser única. Tal limitação é devido ao uso de CPF/CNPJ parcial, onde pode ocorrer casos de pessoas com o mesmo nome completo tenham os mesmo digitos parciais do CPF/CNPJ. Uma forma de diminuir a ocorrência desses casos atípicos é a utilização do estado de registro da pessoa, adicionando uma camada a mais de identificação. Ademais, nós específicos representam PEPs e benefícios (Auxílio Emergencial, Bolsa Família) e relações entre pessoas físicas e pessoas jurídicas (PF–PJ) foram decompostas em arestas granulares (sócio, sócio-representante etc.), conforme esquematizado no Algoritmo 1.

Algorithm 1 Agregação de Dados para Criação de Grafo - Pseudocódigo

AggregateGraphData()
Objetivo: Criar tabela de relações para construção do grafo a partir de múltiplas fontes

1. **Criar tabela temporária** `rede_1` para armazenamento das relações
2. **Inserir relações PJ + razão social** → **PJ + razão social** (empresa sócia de outra empresa)
3. **Inserir relações PF + Nome Completo** → **PJ + razão social** (pessoa física sócia de empresa)
4. **Inserir relações PE + razão social** → **PJ + razão social** (empresa estrangeira sócia de empresa brasileira)
5. **Inserir relações PF + Nome Completo** → **PE + razão social** (representante legal de empresa estrangeira)
6. **Inserir relações PF + Nome Completo** → **PJ + razão social** (representante legal de empresa)
7. **Inserir relações PF + Nome Completo** → **PF + Nome Completo** (representante legal de sócio pessoa física)
8. **Inserir relações PEP** → **PF + Nome Completo** (pessoas expostas politicamente)
9. **Inserir relações BOLSA_FAM** → **PF + Nome Completo** (beneficiários do Bolsa Família)
10. **Inserir relações AUX_EMER** → **PF + Nome Completo** (beneficiários do Auxílio Emergencial)
11. **Criar tabela final** `rede_pep_idx` agrupando relações únicas
12. **Criar índices** para otimização de consultas (`idx_ligacao_origem`, `idx_ligacao_destino`)

Estrutura das relações:
- **Origem:** PEP, BOLSA_FAM, AUX_EMER, PF_[CPF]_[NOME], PJ_[CNPJ], PE_[NOME]
- **Destino:** PF_[CPF]_[NOME], PJ_[CNPJ]
- **Tipo:** descrição da qualificação, ”favorecido”, sigla da função
- **Base:** fonte dos dados (socios, pep, bolsa_familia, aux_emergencial)

A etapa metodológica central combina DCBD com exploração relacional em gra-

fos para detectar padrões e conexões entre PEPs, beneficiários e empresas. O processo é orientado por hipóteses avaliadas iterativamente, permitindo refino de filtros.

A Figura 1 mostra três exemplos de pequenos grafos que são gerados pelo PEP Scan. No grafo (a), por exemplo, o nó central é uma pessoa física que é um prefeito (um PEP), como pode ser visto pela aresta anotada com "PREFEI". Também se observa que essa mesma pessoa física também relaciona-se com uma empresa. Além disso, o grafo também revela que essa pessoa física recebeu auxílio emergencial do governo, o que é destacado pela aresta vermelha.

Em b), tem-se também um PEP que é favorecido do Auxílio Emergencial, porém não relacionado diretamente com qualquer outra entidade. Assim como em a), aresta com atributo "favorecido" indica a relação entre a pessoa física e o programa Auxílio Emergencial, anotado com "AUX_EMER". Por fim, c) evidencia múltiplos papéis societários ("Sócio", "rep-Sócio") conectando a PF a diferentes pessoas jurídicas, destacando a expressividade dos rótulos de aresta e a capacidade de representação de papéis sobrepostos.

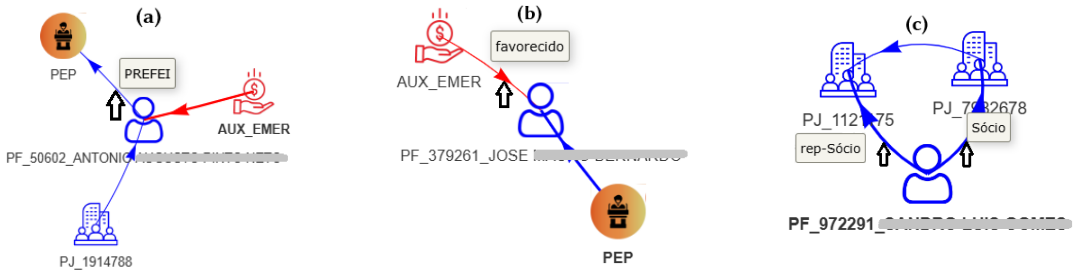

Figura 1. Exemplos de relações nos grafos

Para exploração e explicabilidade, adotou-se a Busca em Largura como mecanismo padrão de navegação e de cálculo do caminho mais curto em grafos não ponderados, com complexidade $O(V+E)$ e controle de profundidade. O pseudocódigo da Busca em Largura está apresentado no Algoritmo 2. A rota `/shortest_path` (Algoritmo 3) implementa o caminho mínimo com restrições: (i) normalização de identificadores (prefixos PF/PJ, truncamento central de CPF/CNPJ e saneamento de acentos); (ii) lista com nós a serem evitados no cálculo por serem atributos ({`PEP, AUX_EMER, BOLSA_FAM`}), prevenindo expansões triviais; (iii) `max_depth` para mitigar explosão combinatória; e (iv) retorno de um subgrafo de visualização serializado em JSON.

Neste contexto, a mineração visual sustenta a análise iterativa por meio de grafos interativos (`pyvis`/`NetworkX`), nos quais filtros e profundidade podem ser ajustados em tempo real. A metodologia de visualização subsequente é empregada para apresentar, revisar e refinar padrões detectados, conforme o diagrama mostrado na Figura 2. Essa abordagem melhora a interpretabilidade e a auditabilidade dos indícios, favorecendo as investigações por analistas e órgãos de controle.

4. PEP Scan: Arquitetura e Funcionalidades

PEP Scan é um protótipo de aplicação web que combina técnicas de mineração de dados, análise de grafos e visualização interativa para identificar possíveis recebimentos indevidos de benefícios por PEPs. Os códigos principais da aplicação e o endereço do site do

Algorithm 2 Busca em Largura - Pseudocódigo

BL(grafo G, nó s)
1. Inicializar fila Q
2. Marcar s como visitado e definir distância$[s] \leftarrow 0$
3. Enfileirar s em Q
4. **while** Q não estiver vazia **do**
   4.1. $u \leftarrow$ desenfileirar(Q)
   4.2. **for** cada vizinho v de u **do**
      4.2.1. **if** v não visitado **then**
         4.2.1.1. Marcar v como visitado
         4.2.1.2. distância$[v] \leftarrow$ distância$[u] + 1$
         4.2.1.3. predecessor$[v] \leftarrow u$
         4.2.1.4. Enfileirar v em Q
      4.2.2. **end if**
   4.3. **end for**
5. **end while**

Algorithm 3 Rota `/shortest_path` via Busca em Largura - Pseudocódigo

ShortestPath(grafo G, nós u, v, conjunto A, profundidade máxima $D_{\max}$)
Entrada: Grafo G, nós u, v, conjunto de nós a evitar A, profundidade máxima $D_{\max}$
Saída: Subgrafo de visualização ou mensagem de erro

1. **if** $u \notin G$ **ou** $v \notin G$ **then**
   1.1. **return** "Entidades não conectadas."
2. **end if**
3. fila $\leftarrow [(u, 0)]$ *(tupla: nó, profundidade)*
4. visitados $\leftarrow \{u\}$
5. predecessor $\leftarrow \emptyset$
6. **while** fila não vazia **do**
   6.1. $(x, d) \leftarrow$ fila.pop(0)
   6.2. **if** $x = v$ **then**
      6.2.1. **break**
   6.3. **end if**
   6.4. **if** $d < D_{\max}$ **then**
      6.4.1. **for** cada y em G.vizinhos(x) **do**
         6.4.1.1. **if** $y \notin$ visitados **e** $y \notin A$ **then**
            6.4.1.1.1. visitados $\leftarrow$ visitados $\cup \{y\}$
            6.4.1.1.2. predecessor$[y] \leftarrow x$
            6.4.1.1.3. fila.append$(y, d + 1)$
         6.4.1.2. **end if**
      6.4.2. **end for**
   6.5. **end if**
7. **end while**
8. path $\leftarrow$ Reconstruir_Caminho(predecessor, u, v)
9. subG $\leftarrow G$.subgraph(path)
10. **return** Preparar_JSON_Visualizacao(subG) *(subgrafo para front-end)*

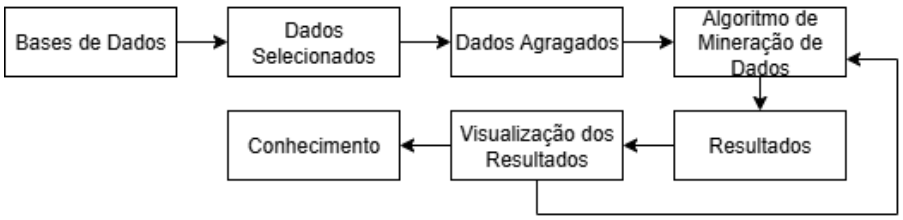

Figura 2. Aplicação de visualização subsequente

protótipo são encontrados no repositório GitHub [5] [6]. O sistema foi concebido para integrar bases públicas heterogêneas, como registros de PEPs, Auxílio Emergencial, Bolsa

[5] `https://github.com/s64864316/PEP_Scan`
[6] `https://pep-scan-tqnn.onrender.com`

Família e CNPJ, e empregar a Busca em Largura para mapear e explorar as conexões complexas. Dessa forma, a ferramenta oferece uma interface intuitiva, onde o uso da MVD permite a exploração interativa das redes de conexões.

A Figura 3 mostra a arquitetura da ferramenta que foi projetada para ser replicável e de baixo custo, em termos de manutenção, promovendo a democratização do acesso a análises complexas. Na parte inferior da Figura 3 observa-se os dados brutos em arquivos CSV, advindos de bases de dados externas. Neste banco de dados, ocorre todo o processo de extração, transformação e carregamento dos dados tratados no mesmo banco, porém em outras tabelas, sendo considerada internamente uma nova camada de dados. Os dados tratados são usados para a geração do grafo bidirecional de relacionamentos, posicionado no lado direito da arquitetura, que é consumido pelo cliente da aplicação, representado na parte superior da Figura. Adicionalmente, a aplicação também consome dados direto do banco, para trazer detalhes adicionais de PEPs, que podem ser exportados pelo usuário.

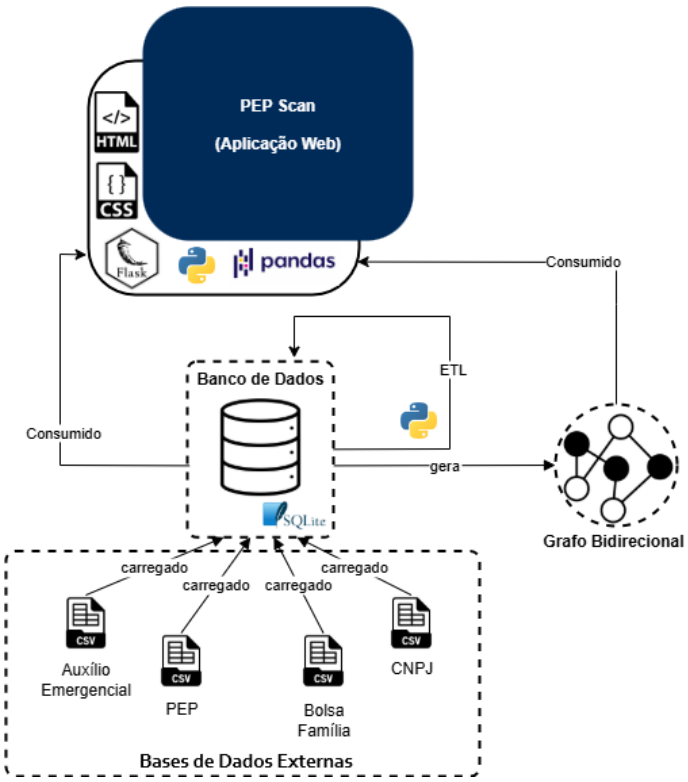

Figura 3. Arquitetura em alto nível do PEP Scan

O método de pesquisa adotado foi o desenvolvimento de um protótipo, seguindo uma abordagem iterativa de engenharia de software e de dados. O alicerce para o projeto foram as bases de dados públicas detalhadas na Tabela 1 foram coletadas e processadas em um *pipeline* de ETL (Extração, Transformação e Carregamento) utilizando Python, Pandas e Polars. O processo de vinculação de registros foi realizado pela concatenação de CPF/CNPJ parciais e nomes, uma limitação que exigiu cautela. A arquitetura do sistema foi baseada em um banco de dados leve (SQLite3) e um servidor web em Python (Flask). O núcleo metodológico foi a aplicação de algoritmos de busca em grafos. Em particular, a Busca em Largura foi adaptado para o domínio do PEP Scan.

Em se tratando de modelagem de dados, a Figura 4 apresenta o Diagrama

Entidade-Relacionamento (DER) que detalha o modelo de dados conceitual para a identificação de anomalias em redes de influência política. Esse modelo é fundamental para a posterior construção do grafo, onde cada entidade se torna um nó e cada relacionamento, uma aresta. No centro do modelo, encontra-se a entidade Pessoa Física (PF), que representa o principal ator na rede. A PF pode se relacionar com outras entidades de diversas maneiras, formando as conexões que serão analisadas. A cardinalidade N nos relacionamentos indica que uma PF pode ter múltiplos vínculos. Portanto, A PF mantém um relacionamento de vínculo público com a entidade PEP com cardinalidade 1 de PEP para PF, indica que uma PEP é, por definição, uma Pessoa Física.

Em um relacionamento vínculo societário, a PF se conecta à entidade Pessoa Jurídica (PJ). A cardinalidade N em ambos os lados (N:N) sugere que uma PF pode ser sócia de várias PJs, e uma PJ pode ter múltiplos sócios PFs. Essa relação é vital para mapear redes de empresas e seus proprietários. Além dos vínculos societários, a PF pode estar ligada a programas sociais. O relacionamento beneficiário conecta a PF ao Bolsa Família, com cardinalidade 1 do Bolsa Família para a PF. Similarmente, o relacionamento favorecido a vincula ao Auxílio Emergencial, também com cardinalidade 1 do Auxílio Emergencial para a PF. A cardinalidade N da PF para ambos os benefícios indica que uma tais programas podem ter vários beneficiários ou favorecidos.

Por fim, o uso de diamantes duplos nos relacionamentos indica que as entidades PEP, Bolsa Femília e Auxílio Emergencial são fracas, ou seja, não podem existir sem a PF. Por outro lado, a geometria da entidade PJ se caracteriza por uma entidade associativa.

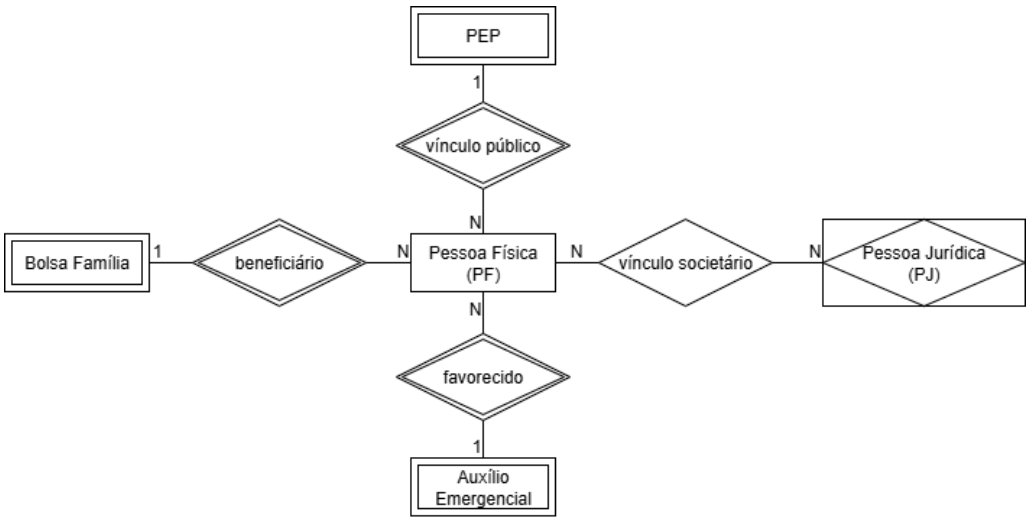

Figura 4. Diagrama entidade-relacionamento - PEP Scan

O PEP Scan foi implementado e funcionalmente validado, apresentando funcionalidades como:

- **Distribuição Geográfica de PEPs:** Na página ”Distribuição de PEPs”, é possível observar como os PEPs estão espalhados geograficamente pelo país, conforme mostrado na Figura 5.
- **Análise Direta de Auxílios:** A ferramenta identificou um número significativo de PEPs que receberam auxílios governamentais. A Tabela 7 sumariza esses achados para o Auxílio Emergencial e o Bolsa Família.

- **Detecção de Relações Indiretas:** A funcionalidade ”Caminho Mais Curto”e ”Relacionamentos”demonstrou a capacidade de encontrar conexões indiretas. Por exemplo, foram identificados casos suspeitos de profundidade 3, onde um beneficiário estava ligado a três PEPs distintos através de uma mesma empresa, conforme ilustrado na Figura 6.
- **Visualização Interativa:** A aplicação gerou grafos interativos que permitem a exploração dinâmica de relacionamentos entre PEPs, beneficiários e empresas, revelando a dimensão e a complexidade das redes.

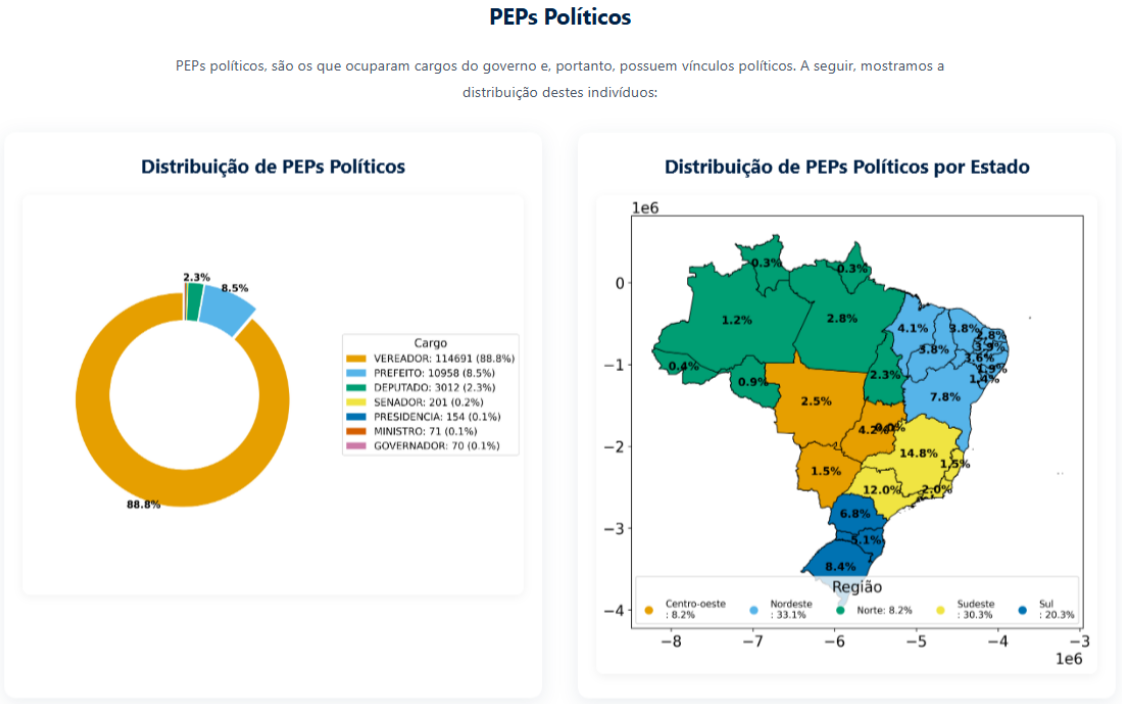

Figura 5. Distribuição de PEPs políticos

Tabela 7. Quantidade de PEPs que receberam auxílios governamentais

Função	Quantidade
Vereador	422
Prefeito	20
Senador	1

(a) Auxílio Emergencial

Função	Quantidade
Vereador	2

(b) Bolsa Família

5. Avaliações

O protótipo desenvolvido foi avaliado considerando dois pontos principais: qualidade da arquitetura proposta e usabilidade.

5.1. Avaliação da Qualidade Arquitetural

Esta seção apresenta alguns dados sobre a qualidade arquitetural do PEP Scan. A ferramenta usada na análise foi a Sonar Graph Architect [von Zitzewitz 2019], uma ferramenta de análise estática que possui o cálculo de diversas métricas de qualidade e permite

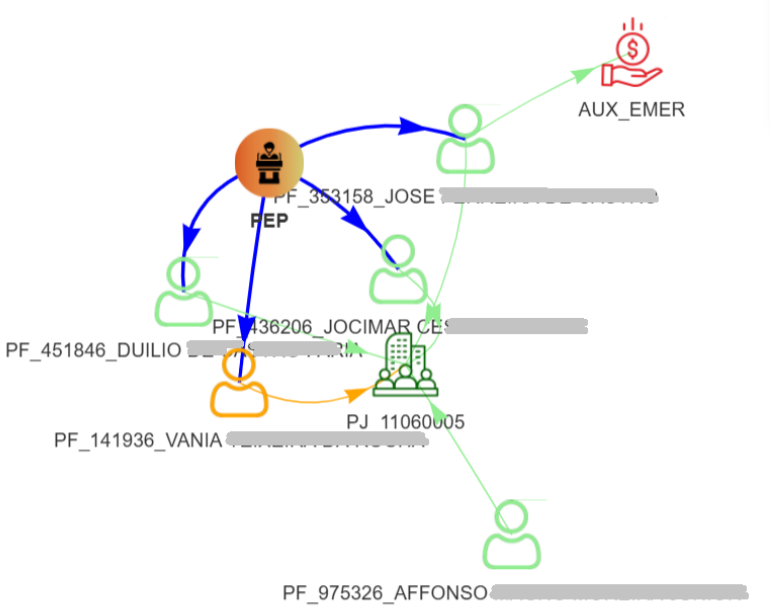

Figura 6. Exemplo de conexões suspeitas

a geração de vários tipos de grafos de dependência. Devido à limitação de espaço, é mostrado aqui alguns pontos de destaque da arquitetura.

A Figura 6 mostra uma representação gráfica chamada *Tree Map* que representa a arquitetura do sistema. As camadas inferiores (em cinza) representam a árvore de diretórios do sistema. As representações gráficas que se assemelham a prédios representam os arquivos/modulos Python. Nesta figura, a largura representa o tamanho do arquivo em linhas de código, a cor representa a complexidade média e a altura é a profundidade de aninhamento dos blocos.

Note-se que em geral, os três indicadores possuem uma harmonia para 12 dos 13 indicadores. Apenas um dos módulos se destaca na altura, pela profundidade de aninhamento dos blocos, e na cor, pela complexidade média que ele possui. Este módulo é *copy.py* e possui 610 linhas de código.

No entanto, mesmo com um dos módulos apresentando divergência com os demais, as principais métricas de qualidade para o sistema possuem valores muito bons, como pode ser visto na Tabela 8. A complexidade ciclomática (CC) mede basicamente o número de fluxos alternativos dentro de métodos, sendo também um parâmetro para o número de casos de teste que precisam ser criados. O valor médio para o sistema todo é 1,29, sendo bem baixo. Os únicos métodos que possuem complexidade ciclomática mais alta são os responsáveis pelo cálculo do caminho mais curto, sendo eles: o método *shortest_path()* com CC = 7 e o método *generate_graph()* com CC = 6.

Também nota-se na tabela que a implementação do sistema não apresenta ciclos, o que é considerado bastante prejudicial para a manutenibilidade. Assim, o nível de manutenibilidade do sistema está em 100%. Outra métrica interessante é a LCOM4, que mede falta de coesão em classes. Todas as classes do sistema possuem coesão 1, que é o nível mais alto, mostrando que todas as classes possuem uma responsabilidade

única. Somente uma delas possui LCOM com valor 3, que é a classe *GraphManager*. Analisando essa classe mais profundamente, nota-se que ela poderia ser refatorada de forma a melhorar sua coesão e, consequentemente, a métrica LCOM4.

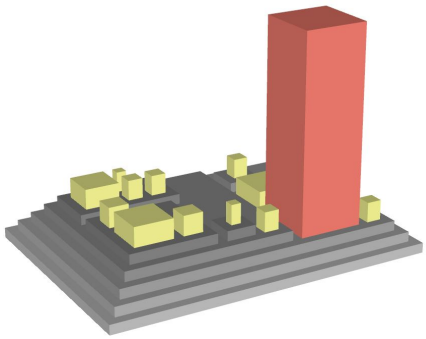

Figura 7. Tree Map da arquitetura do PEP Scan

Tabela 8. Métricas de Qualidade do PEP Scan

Métrica	Valor
Complexidade Ciclomática Média	1,29
Número de pacotes Python em ciclo	0
Nível de manutenibilidade	100%
LCOM 4	Maior valor = 3 *Classe: Graph Manager*

5.2. Avaliação da Usabilidade

A usabilidade foi avaliada por inspeção com base nas 10 heurísticas de Jakob Nielsen [Nielsen 1994a, Nielsen 1994b], adaptadas ao contexto de exploração de grafos em dados públicos. A avaliação foi conduzida via formulário digital e executada por seis profisionais, sendo três deles especialistas em usabilidade e interface do usuário e outros três da área de tecnologia da informação. Tal avaliação visou identificar problemas de visibilidade de estado, consistência, prevenção/recuperação de erros e eficiência de uso (atalhos e filtros). Dentre cada uma das heurísticas avaliadas pelos respondentes, foram selecionadas para melhorias aquelas que apresentaram pelo menos uma avaliação de Severidade 3 (Problema Grave) ou Severidade 4 (Problema Catastrófico), com o intuito de priorizar o aperfeiçoamento das heurísticas mais relevantes.

É importante destacar que a avaliação heurística aplicada neste projeto não envolveu avaliar, separadamente, um ou mais fluxos de trabalho, como por exemplo, o julgamento de uma ação específica previamente definida. O foco recaiu sobre os princípios gerais de interação e descobribilidade de funcionalidades.

Tal metodologia revelou problemas críticos de usabilidade. A heurística ”Prevenção de Erros”apresentou uma severidade catastrófica, indicando falhas na validação de inputs. As heurísticas ”Reconhecimento em vez de Memorização”e ”Flexibilidade e Eficiência”também apresentaram problemas graves, sugerindo uma alta carga cognitiva e a falta de atalhos ou personalização para usuários experientes. As correções

aplicadas, como validação em tempo real e melhoria da hierarquia visual, foram cruciais para aprimorar a experiência do usuário.

Adicionalmente, a heurística de Prevenção de Erros focou em projetar o sistema de forma a evitar que o usuário cometa equívocos em primeiro lugar, em vez de apenas oferecer mensagens de erro após o ocorrido. Isso pode ser alcançado através de validações em tempo real, sugestões contextuais, confirmações antes de ações destrutivas ou a remoção de opções que levariam a um estado inválido. No PEP Scan, dada a natureza crítica dos dados e análises, a prevenção de erros é fundamental. Isso inclui validações em campos de busca, como CPF e CNPJ, para evitar consultas mal-formatadas. A prevenção de erros minimiza a frustração do usuário e o tempo gasto na correção de equívocos.

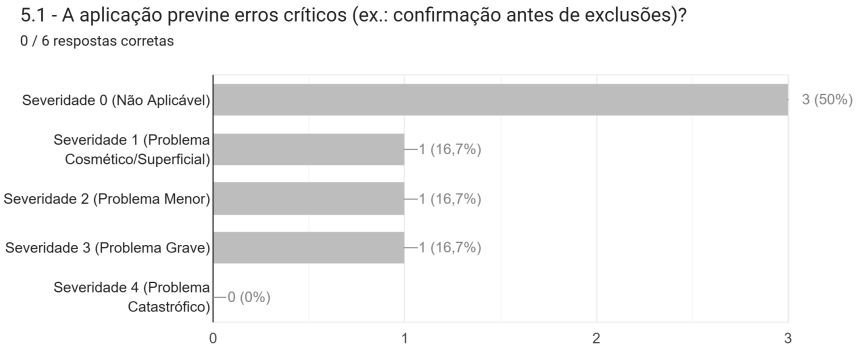

Figura 8. Heurística de usabilidade - Avaliação sobre prevenção de erros críticos

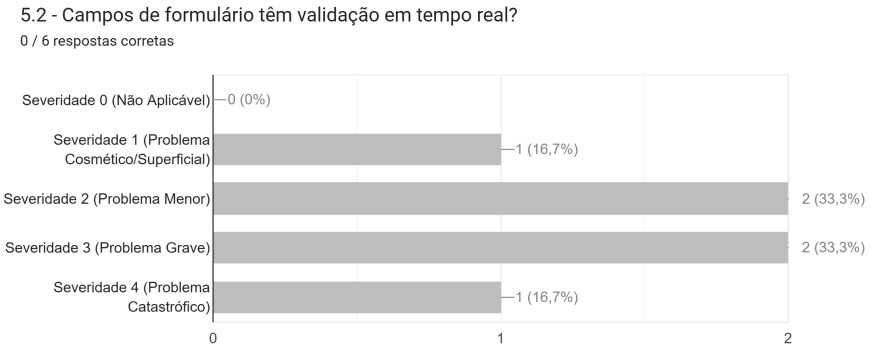

Figura 9. Heurística de usabilidade - Avaliação sobre validação em inputs em tempo real

Essa heurística foi considerada uma das mais críticas durante a avaliação. A avaliação 5.2, apresentada na Figura 9 apresenta uma severidade 4 (catastrófico) e dois problemas de severidade grave, indicando uma falha significativa na validação em tempo real de campos de formulário. Isso pode levar a submissões de dados incorretos, interrupções no fluxo de trabalho e frustração considerável.

De forma similar, a avaliação 5.1, dada pela Figura 8 revela um problema de severidade grave quanto à prevenção de erros críticos, o que representa um risco considerável

de perda de dados ou ações irreversíveis. A falta de mecanismos proativos para evitar erros impõe uma carga desnecessária ao usuário para ser cauteloso e verificar suas entradas, impactando negativamente a eficiência e a segurança da aplicação.

Portanto, com base nos resultados da avaliação, foram implementadas melhorias chave para aprimorar a prevenção de erros e a segurança do PEP Scan. As principais correções foram focadas na validação de entrada de dados e na interação do usuário. As melhorias aplicadas na página "Caminho Mais Curto"foram replicadas na página "Relacionamentos". A Figura 10 apresenta um exemplo de melhoria da heurística de Prevenção de Erros, validando as entradas de usuários na ferramenta.

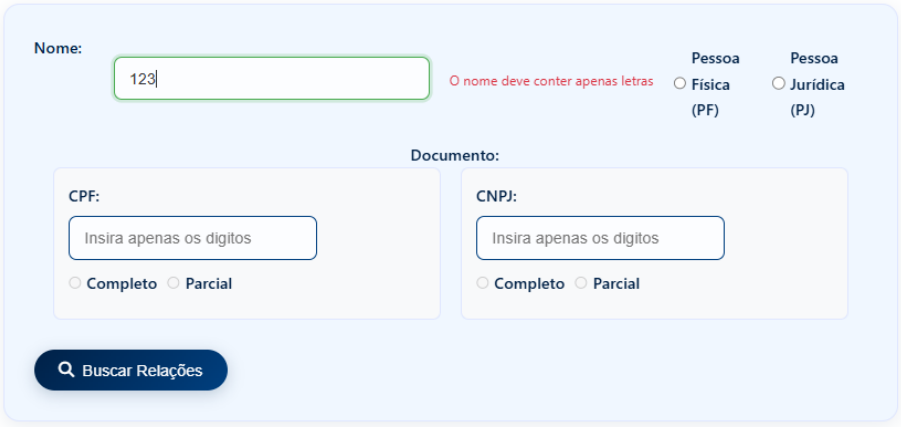

Figura 10. Relacionamentos - versão final - com validações de entradas

6. Trabalhos Relacionados

A detecção de atividades fraudulentas é um campo de pesquisa consolidado, com uma vasta literatura que abrange diversas técnicas, desde abordagens estatísticas tradicionais até métodos avançados de mineração de dados e aprendizado de máquina. Neste cenário, com a crescente digitalização e interconexão de dados, as abordagens baseadas em grafos têm se destacado como uma técnica particularmente promissora. Esses métodos, conhecidos como DABG, são projetados para analisar padrões de conectividade em redes e identificar comportamentos anômalos que, de outra forma, ficariam ocultos na complexidade dos dados.

Nesse contexto, uma revisão sistemática da literatura conduzida por [Pourhabibi et al. 2020] analisou o cenário de pesquisa em DABG aplicado à detecção de fraudes, examinando 39 artigos publicados entre 2007 e 2018. O estudo categorizou as pesquisas em um arcabouço de classificação robusto, considerando a disponibilidade de rótulos de dados (supervisionado, não supervisionado, semi-supervisionado), a natureza da rede de entrada, os tipos de anomalias detectadas e os métodos de grafos empregados. A análise de Pourhabibi et al. revelou que a maioria das pesquisas (87.2%) se concentra em técnicas de aprendizado não supervisionado, dada a escassez de dados rotulados no mundo real. A revisão também mostrou um uso predominante de dados de redes sociais online (OSNs), com aplicações em setores tradicionais, como seguros, telecomunicações e finanças (bancos, transações de varejo, *trading*, e anti-lavagem de dinheiro). Esses estudos se concentraram em identificar fraudes em cenários como reclamações de seguro, movimentações financeiras e manipulação de mercado.

Em termos de tipologia de anomalias, Pourhabibi et al. notaram que a maioria dos estudos foca na identificação de subgrafos anômalos, dada a natureza coletiva das fraudes em rede. Os métodos mais utilizados para esse fim foram as abordagens baseadas em comunidade (35.9%) e probabilísticas (25.6%). A revisão também destacou desafios significativos na área, como a falta de dados públicos, a necessidade de lidar com a natureza dinâmica e *multiplex* (múltiplas interações) das redes, e a dependência de intervenção humana na extração de características (engenharia de *features*). Como recomendações, os autores sugerem o uso de dados sintéticos mais realistas, a adaptação de algoritmos para redes dinâmicas e multiplex, e o uso de técnicas de aprendizado de representação de grafos para automatizar a extração de características.

Considerando-se o cenário apresentado, uma grande lacuna surge. Embora a pesquisa de Pourhabibi et al. ofereça uma visão abrangente sobre o estado da arte na detecção de fraudes em diversos setores, ela não aborda um domínio de aplicação que consideramos fundamental para a ciência de dados no setor público: o contexto político e de combate à corrupção. Ao examinar a distribuição dos trabalhos revisados, é notável a ausência de estudos que aplicam o DABG para analisar padrões de comportamento de Pessoas Politicamente Expostas ou para detectar fraudes em redes de influência política. A totalidade dos casos de uso se restringe a domínios comerciais e financeiros, como seguros e sistemas bancários. Essa lacuna é particularmente relevante, pois a corrupção e o mau uso de recursos públicos frequentemente se manifestam por meio de conexões complexas e ocultas entre indivíduos e entidades, que são precisamente o tipo de padrão que as técnicas de DABG são projetadas para revelar.

Desta maneira, pesquisa que resultou no PEP Scan busca preencher tal lacuna, adaptando e estendendo a aplicação de técnicas de mineração de grafos para um domínio de interesse público. Diferentemente dos trabalhos revisados por Pourhabibi et al., que se concentram em detectar fraudes em dados de empresas ou redes sociais, o presente trabalho direciona o foco para o rastreamento de uso indevido de recursos públicos por parte de agentes políticos. Para tanto, utilizou-se uma metodologia inspirada na análise de grafos para, em vez de focar em uma aplicação privada de detecção de fraude, mapear as interconexões complexas entre dados públicos e identificar padrões de corrupção que seriam invisíveis em análises tradicionais.

Ademais, outro importante trabalho relacionado é a Operação Serenata de Amor (OSA) [Irio Musskopf 2025], um projeto de tecnologia cívica brasileiro. Destaca-se como um marco na utilização de dados abertos governamentais para o controle social e o combate à corrupção no Brasil. O projeto foi idealizado por cidadãos motivados a auditar as contas públicas e fiscalizar os reembolsos de despesas dos parlamentares. A OSA se baseia em dados abertos governamentais disponibilizados pelo governo brasileiro, um ambiente legal que, desde 2011, com a criação do Portal Brasileiro de Dados Abertos e a Lei de Acesso à Informação (LAI), busca maior transparência. O projeto ”Rosie”, um robô que usa algoritmos de aprendizado de máquina não supervisionado, foi desenvolvido para analisar notas fiscais e recibos de gastos dos deputados, procurando por irregularidades e estimando a probabilidade de corrupção. O sistema de visualização de dados ”Jarbas”foi criado para facilitar a consulta das informações geradas pela ”Rosie”para que o cidadão pudesse validar as suspeitas e, se necessário, denunciá-las.

Além disso, o projeto demonstrou que disponibilizar dados é insuficiente; a

informação deve ser acessível ao cidadão. A OSA converteu dados densos (exigindo 8 GB de RAM) em arquivos simplificados, gerando valor público ao facilitar o controle social, apesar do baixo engajamento histórico no Brasil. Através da "Rosie", identificou gastos suspeitos e estimulou denúncias diretas em mídias sociais. Contudo, a OSA enfrenta desafios como a dependência de um ambiente legal favorável e a necessidade de dados governamentais utilizáveis. Outrossim, por focar em uma única categoria de despesa, não abrange a complexidade e heterogeneidade dos dados que caracterizam os PEPs.

Nesse sentido, o projeto PEP Scan compartilha da mesma inspiração do modelo de sucesso da Operação Serenata de Amor, mas se diferencia e inova em sua abordagem e escopo. Enquanto a OSA foca na fiscalização de um tipo específico de reembolso parlamentar, o PEP Scan adota uma visão mais ampla e complexa, direcionando-se para o monitoramento de PEPs e suas redes de relacionamento. Isso significa ir além de uma única fonte de dados e integrar informações de múltiplas bases de dados governamentais heterogêneas, como registros de benefícios sociais, e registros de empresas, para mapear conexões que podem indicar fraudes. Ao contrário da abordagem da OSA que se concentra na detecção de despesas suspeitas individuais, o PEP Scan busca identificar padrões de comportamento atípicos e conexões complexas que envolvem não apenas a PEP, mas também suas conexões. Para isso, utilizou-se a Busca em Largura para mapear essas conexões e a MVD para permitir a exploração interativa das redes, o que se alinha com as recomendações de literaturas que destacam a importância de abordar a natureza *multiplex* das interações. O PEP Scan não apenas automatiza o processo de análise de dados, mas também oferece uma interface que capacita o cidadão a "dar significado"aos dados de forma visual e intuitiva, superando a barreira de acessibilidade apontada tanto por ativistas quanto por acadêmicos no contexto da OSA.

Por fim, outro ponto de complementação é que, ao passo que a OSA também opera com um modelo deciência cidadã, onde os cidadãos voluntários participam da fiscalização de forma colaborativa, o PEP Scan adota uma abordagem complementar, focando em fornecer uma ferramenta técnica robusta e replicável que pode ser utilizada tanto por cidadãos, quanto por cientistas de dados, jornalistas investigativos e órgãos de controle, agindo como um catalisador para a análise aprofundada de corrupção sistêmica. Ao documentar e tornar público o processo técnico e a metodologia, a presente pesquisa contribui para a consolidação de novas formas de produção de conhecimento, envolvendo ciência e cidadania, conforme o espírito da própria Operação Serenata de Amor. Dessa forma, o PEP Scan avança no campo de tecnologias cívicas ao transcender a fiscalização pontual de gastos e propor um modelo de análise de redes complexas para combater a corrupção de forma mais estrutural.

7. Conclusões e Relacionamento com Desafios da Área de SI

Este trabalho oferece diversas contribuições relevantes para a área de Sistemas de Informação e para a sociedade. Primeiramente, estabelece um modelo metodológico replicável, apresentando uma metodologia para a integração e análise de dados públicos heterogêneos que combina DCBD e MVD para abordar o problema de forma holística em contextos de políticas públicas.

Em termos de aplicação, o PEP Scan consolida-se como uma ferramenta prática de

combate à corrupção, alavancando a tecnologia para fortalecer a *accountability* e a transparência. Ele fornece um protótipo funcional para órgãos de controle, jornalistas investigativos e cidadãos, transformando dados brutos em inteligência acionável. Simultaneamente, a pesquisa demonstra inovação em visualização de dados ao comprovar a eficácia da aplicação de bibliotecas de visualização web-nativas de baixo custo de manutenção para democratizar o acesso à análise de redes complexas, superando as limitações de ferramentas *desktop* e promovendo a "ciência cidadã".

Adicionalmente, houve um forte foco na usabilidade em sistemas de análise. A inclusão de uma avaliação heurística rigorosa demonstrou a importância de a usabilidade ser uma preocupação central no desenvolvimento de sistemas tecnicamente complexos. A identificação de deficiências críticas e a implementação de melhorias garantem que a ferramenta não apenas seja robusta, mas também eficaz na prática.

Ademais, o PEP Scan é um passo inicial, mas chave, na demonstração de que a ciência de dados pode ser uma aliada estratégica no enfrentamento da corrupção. O trabalho evidenciou que a transparência é um valor alcançável através da colaboração entre academia, governo e sociedade, pavimentando o caminho para futuras iniciativas audaciosas.

Por fim, este trabalho se relaciona diretamente com as temáticas dos grandes desafios de pesquisa em SI no Brasil. A ferramenta apresentada contribui para a democracia virtual, pois permite que cidadãos possam fazer consultas sobre PEPs para compreender a relação desses com outros agentes políticos e empresas. Além disso, a ferramenta está intimamente ligada ao tema Metodologias e Técnicas para Participação dos Cidadãos, já que claramente permite o envolvimento desses no cotidiano da política nacional, disseminando conhecimento que é de direito da população.

Referências

Aggarwal, C. C. et al. (2015). *Data mining: the textbook*, volume 1. Springer.

Akoglu, L., Tong, H., and Koutra, D. (2015). Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery*, 29(3):626–688.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, volume 3, pages 361–362.

Borgatti, S. P., Mehra, A., Brass, D. J., and Labianca, G. (2009). Network analysis in the social sciences. *science*, 323(5916):892–895.

Brasil (2017). Resolução nº 29, de 7 de dezembro de 2017. Dispõe sobre os procedimentos a serem observados pela União, pelos Estados, pelo Distrito Federal e pelos Municípios para a identificação, a qualificação e o cadastramento de beneficiários finais de pessoas jurídicas e outras entidades sob sua administração direta e indireta.

Fayyad, U. (1997). Knowledge discovery in databases: An overview. In *International Conference on Inductive Logic Programming*, pages 1–16. Springer.

Financial Action Task Force (FATF) (2013). Guidance: Politically exposed persons (recommendations 12 and 22). `https://www.fatf-gafi.org/`. Accesso em: 30 jun. 2025.

GCU (2023). Pep. Disponível em: https://portaldatransparencia.gov.br/download-de-dados/pep. Acesso em: 12 ago. 2023.

Han, J., Pei, J., and Tong, H. (2022). *Data mining: concepts and techniques*. Morgan kaufmann.

Irio Musskopf (2025). Operação serenata de amor. Disponível em: https://serenata.ai/. Acesso em: 4 fev. 2025.

Keim, D. A. and Kriegel, H.-P. (2002). Visualization techniques for mining large databases: A comparison. *IEEE Transactions on knowledge and data engineering*, 8(6):923–938.

Kerche, F. and Marona, M. (2022). *A política no banco dos réus: a Operação Lava Jato e a erosão da democracia no Brasil*. Autêntica Editora.

King, C. and Walker, C. (2016). *Dirty assets: Emerging issues in the regulation of criminal and terrorist assets*. Routledge.

Newman, M. E. J. (2018). *Networks: An Introduction*. Oxford University Press.

Nielsen, J. (1994a). *Usability engineering*. Morgan Kaufmann.

Nielsen, J. (1994b). Usability inspection methods. In *Conference companion on Human factors in computing systems*, pages 413–414.

Pourhabibi, T., Ong, K.-L., Kam, B. H., and Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, 133:113303.

Sharman, J. C. (2009). Politically exposed persons: from victims to villains to victors. *Crime, Law and Social Change*, 52(1):67–87.

Transparência Internacional (2024). Índice de percepção da corrupção 2023. https://transparenciainternacional.org.br/publicacoes/indice-de-percepcao-da-corrupcao-2023/. Acesso em: 16 dez. 2024.

von Zitzewitz, A. (2019). Mitigating technical and architectural debt with sonargraph. In *2019 IEEE/ACM International Conference on Technical Debt (TechDebt)*, pages 66–67. IEEE.

West, D. B. (2001). Introduction to graph theory.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., and Data, M. (2005). Practical machine learning tools and techniques. In *Data mining*, volume 2, pages 403–413. Elsevier Amsterdam, The Netherlands.