Uso de banco de dados orientado a grafos na detecção de fraudes nas cotas para exercício da atividade parlamentar

1

Abstract. *This paper proposes the use of graph oriented databases to detect possible frauds in Quota for the Exercise of Parliamentary Activity, and the relationships between CEAP and open data from TSE regarding donations to 2014 brazilian elections. The use of these technologies, facilitates the manipulation of closely related data, both in terms of query complexity, and information visualization. The proposal in question was validated with a case study, using the open data of the Quota for the Exercise of the Parliamentary Activity of the Chamber of Deputies, and open data from TSE. It was developed an ETL to extract the data and fill the database, then the queries were made to detect the possible frauds and to obtain information about the data.*

Resumo. *Este artigo propõe o uso da tecnologia de banco de dados orientado a grafos para a detecção de possíveis fraudes na Cota para o Exercício da Atividade Parlamentar (CEAP), e o relacionamento da CEAP com dados abertos do TSE referentes as doações nas eleições de 2014. O uso dessas tecnologias facilita a manipulação de dados relacionados entre si, tanto em questão de complexidade na consulta, quanto em relação a visualização da informação. A proposta em questão foi validada com um estudo de caso, utilizando os dados abertos da Cota para o Exercício da Atividade Parlamentar da Câmara dos Deputados, e os dados abertos fornecidos pelo TSE. Foi desenvolvido um ETL para extrair os dados e popular o banco de dados, em seguida as consultas foram realizadas para detectar as possíveis fraudes e obter informações a respeito dos dados.*

1. Introdução

A política de transparência no Brasil surgiu há alguns anos em nossa sociedade, e tem como principal objetivo auxiliar na confiança da população sobre os serviços prestados pelo governo. O Brasil recentemente passou por um grande caso de corrupção, que foi a lava jato, e políticas de transparência também tem suas vertentes no combate a corrupção como mostra [Diirr and Cappelli]. O estudo feito por [Abramo 2000] compara as relações entre índices de percepção de corrupção e outros indicadores de alguns países latino americanos, e o Brasil se encontra na quadragésima nona posição em um ranking de corrupção dentre 90 países. Já o estudo feito por [Filgueiras 2009] analisa uma pesquisa de opinião, na qual a Câmara dos Vereadores e a Câmara dos Deputados são as instituições com maior percepção de corrupção.

A análise feita por [Filgueiras 2009], direcionou a escolha do conjunto de dados utilizado neste trabalho, que foi a Cota para o Exercício da Atividade Parlamentar (antiga verba indenizatória). É uma cota única mensal destinada a custear os gastos dos deputados

exclusivamente vinculados ao exercício da atividade parlamentar. O Ato da Mesa número 43 de 2009, detalha as regras para o uso da CEAP, entretanto um deputado pode realizar algumas transações que não são observadas facilmente pelos responsáveis em fiscalizar essas transações. Por exemplo, o artigo 4, parágrafo 13 do Ato da Mesa número 43 de 2009, diz: *"Não se admitirá a utilização da Cota para ressarcimento de despesas relativas a bens fornecidos ou serviços prestados por empresa ou entidade da qual o proprietário ou detentor de qualquer participação seja o Deputado ou parente seu até terceiro grau."* Dessa forma, o Deputado pode realizar transações que violam essa regra, sendo inviável verificar as relações de parentesco de cada Deputado em cada transação, especialmente se utilizarem tecnologias inadequadas.

Portanto, justifica-se o uso de um banco de dados orientado a grafo para identificar os relacionamentos envolvendo cada transação de um Deputado. Um banco de dados relacional também consegue resolver esse problema, entretanto, com um custo e complexidade bem maior em relação a um banco de dados orientado a grafo. Isso se deve porque os relacionamentos são evidenciados na estrutura de um grafo de forma muito mais natural e simples, onde cada entidade é representada como um nó do grafo e se relaciona com outras entidades por meio de arestas. Devido a essas particularidades, os bancos de dados em grafo vem ganhando bastante popularidade ultimamente, tanto em pesquisas científicas, quanto em uso comercial, como mostra os trabalhos feitos por [Barmpis and Kolovos 2014] e [Labute and Dombroski 2014]. Já o uso desse tipo de banco de dados, como mostra o estudo feito por [Nayak et al. 2013], varia de aplicações para redes sociais, bioinformática, software de recomendação e etc.

O objetivo geral deste trabalho é implementar um banco de dados baseado em grafo para evidenciar relacionamentos nas transações dos Deputados que violam o artigo 4, parágrafo 13 do Ato da Mesa número 43 de 2009, que regula a CEAP, e realizar o cruzamento com os dados abertos de doações das eleições de 2014, em busca de relacionamentos entre parlamentares e empresas. Para a implementação deste banco de dados, foi utilizado o Sistema Gerenciador de Banco de Dados (SGBD) NoSQL OrientDB. Para validar o modelo do banco de dados, algumas consultas de vínculos entre os deputados e as empresas são apresentados.

Este artigo está dividido nas seguintes seções: A Seção 2 apresenta o referencial teórico, abordando o tema de banco de dados orientado a grafos; Na Seção 3 é apresentado em detalhes o desenvolvimento do projeto, e por fim, a Seção 4 apresenta as conclusões deste trabalho.

2. Banco de dados orientado a grafos

2.1 Definição de um grafo

O primeiro passo para entender um SGBD orientado a grafos, é entender a estrutura de um grafo. A definição formal de um grafo pode ser feita da seguinte forma: Um grafo G é uma tripla ordenada $(V(G), E(G), \psi g)$, que consiste de um conjunto não vazio $V(G)$ de vértices, um conjunto $E(G)$, disjunto do conjunto $V(G)$, de arestas, e uma função de incidência ψg que associa cada aresta de G um par não ordenado (não necessáriamente distinto) de vértices de G. Dessa forma, se e é uma aresta e u e v são vértices, de tal modo que $\psi g(e) = uv$, então, diz-se que e faz a união de u e v; Os vértices u e v são chamados de extremidades de e [Bondy et al. 1976].

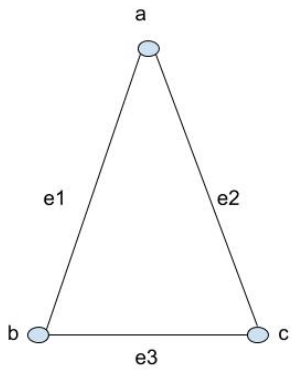

Figure 1. Exemplo de uma estrutura de grafo

A Figura 1 apresenta um grafo, em que o conjunto $V(G)$ de vértices é não vazio e composto por três vértices, $V(G) = \{a, b, c\}$. Já o conjunto $E(G)$ de arestas é composto por três arestas, $E(G) = \{e1, e2, e3\}$. De forma que a função de incidência ψg é definida da seguinte maneira: $\psi g(e1) = ab$, $\psi g(e2) = ac$ e $\psi g(e3) = bc$. Portanto, a definição formal do grafo acima é $(\{a, b, c\}, \{e1, e2, e3\}, \psi g(e1) = ab, \psi g(e2) = ac\, \psi g(e3) = bc)$.

2.2 SGBD NoSQL orientado a grafos

Os SGBD NoSQL orientado a grafos armazenam os dados em uma estrutura de grafo. Resumidamente, um grafo é um conjunto de nós e arestas, em que os nós são os objetos e as arestas representam o relacionamente entre dois objetos (nós). Esses SGBD usam a técnica conhecida como *index free adjacency*, em que cada nó possui um ponteiro diretamente para nós adjacentes. Essa técnica permite que uma grande quantidade de nós seja percorrida de forma eficiente [Nayak et al. 2013].

As consultas são feitas seguindo a ideia de percorrimento do grafo, o que torna esses SGBD mais eficientes que SGBD relacionais. O principal ponto de seu uso é em dados que são bastante relacionados entre si, pois a estrutura de um grafo expõe naturalmente os relacionamentos entre os objetos. O maior representante dessa categoria de SGBD é o Neo4j, que em sua versão livre não suporta a distribuição dos dados. Por esse motivo o OrientDB foi escolhido para a implementação do banco de dados deste artigo.

2.3 OrientDB

O OrientDB, é um SGBD de código aberto sob a licensa Apache. Ele é o primeiro SGBD NoSQL multi modelo, com suporte a uma arquitetura distribuída e orientado a grafos. Sendo assim o OrientDB suporta operações com documentos, chave/valor e grafos. Essa característica garante flexibilidade para manipular os dados dentro do OrientDB, sendo possível armazenar os dados tanto como grafos ou como documentos no mesmo banco de dados.

O OrientDB é implementado utilizando a linguagem java, tendo sua primeira versão disponível no ano de 2010. Ele possui alta flexibilidade para definir o esquema do banco de dados, podendo ser *Schema-free*, *Schema-hybrid* ou *Schema-full*. A sua linguagem de consulta é derivada do SQL o que é bastante vantajoso para aqueles que possuem experiência com bancos de dados relacionais, e além disso ele utiliza o modelo de transações ACID que é algo mais comum no grupo dos SGBD relacionais, isso demonstra que o OrientDB presa pela integridade dos dados ao mesmo tempo que também fornece um suporte a particionamento dos dados.

Todas essas características fazem com que o OrientDB seja um SGBD bastante flexível e confiável para se utilizar em diversas aplicações. As operações utilizando grafos em específico, vem ganhando bastante visibilidade pois funciona muito bem em certos domínios de aplicação. Como foi mencionado em [Gelbmann 2014] a popularidade dos SGBD orientados a grafos vem crescendo bastante nos últimos anos, e essas características ajudam a explicar o porque de banco de dados como o OrientDB e Neo4j estarem sendo utilizados em tantas aplicações.

3. Desenvolvimento

Esta seção descreve o processo de desenvolvimento desta pesquisa. Inicialmente, foram obtidos os dados abertos da CEAP no site da Câmara dos Deputados, e as receitas dos candidatos a deputado federal nas eleições de 2014 no site do TSE. Em seguida, foram feitas algumas transformações nos dados antes da fase de carregamento para o OrientDB. Foi desenvolvido um modelo de dados seguindo a modelagem proposta no trabalho de [Van Erven et al. 2018], que representa os dados persistidos no OrientDB. Em seguida, foram desenvolvidas as consultas para obter os relacionamentos entre os dados da CEAP e os dados das receitas nas eleições de 2014.

3.1 Dados Abertos

Foram utilizadas duas bases de dados abertos para o desenvolvimento deste trabalho. A primeira é referente aos dados da cota para exercício da atividade parlamentar, e pode ser obtida no seguinte site da Câmara dos Deputados[1]. A segunda base diz respeito as doações que cada deputado recebeu de empresas ou pessoas físicas, para a campanha eleitoral de 2014, e pode ser obtida no seguinte site do Tribunal Superior Eleitoral [2].

Como foi mencionado na Seção 1, a iniciativa de disponibilizar dados e informações por meio de portais tem como um dos objetivos melhorar a confiança da população nos serviços prestados pelo governo. A transparência governamental é , portanto, uma ótima iniciativa e extremamente benéfica para a população. Porém ainda existem diversos pontos a serem melhorados para que estudos sejam feitos de forma mais rápida e precisa. Um dos maiores desafios desse projeto foi trabalhar com as bases de dados abertas mencionadas nessa seção.

Trabalhar com esses dados abertos se tornou um desafio, porque, cada instituição fornece os dados da própria maneira, essa falta de padronização dificultou bastante a fase

[1]http://www2.camara.leg.br/transparencia/cota-para-exercicio-da-atividade-parlamentar/dados-abertos-cota-parlamentar

[2]http://www.tse.jus.br/eleitor-e-eleicoes/estatisticas/repositorio-de-dados-eleitorais-1/repositorio-de-dados-eleitorais

de carregamento dos dados. Isso se deve ao fato de que a base de dados do TSE fornece o cpf de cada candidato, já a base da CEAP não fornece um identificador único para o deputado. Portanto, no momento de realizar o cruzamento não existia um identificador único em ambas as bases para fazer o relacionamento entre as bases de dados. Para resolver esse problema, se fez necessário realizar o cruzamento entre as bases por meio do nome de cada candidato, o que também foi um desafio, já que cada base utiliza um nome diferente para cada candidato. Os detalhes do processo de extração, transformação e carregamento serão fornecidos na Seção 3.3, o ponto principal a ser levantado é que trabalhar com dados abertos no Brasil, para realizar estudos e pesquisas, pode se tornar um desafio devido a falta de padronização entre as bases de dados.

3.2 Modelo de Dados

A modelagem dos dados seguiu o modelo GRAPHED [Van Erven et al. 2018]. O trabalho em questão, busca propor formas de modelagem dos dados para bancos de dados orientado a grafos, uma área já desenvolvida no universo dos SGBD relacionais, mas ainda em evolução na categoria de SGBD orientado a grafos.

A modelagem tem como objetivo dar uma visão geral de como os dados estão organizados no banco de dados, suas propriedades e relacionamentos. Dessa forma, o modelo de dados desenvolvido busca evidenciar as propriedades dos parlamentares, tais como: nome, partido e unidade federativa. Além disso apresenta características que identificam uma empresa como nome e CNPJ, e características atreladas a transação que o deputado faz com uma empresa como o valor e descrição da transação. A Figura 2 a seguir apresenta a modelagem desenvolvida.

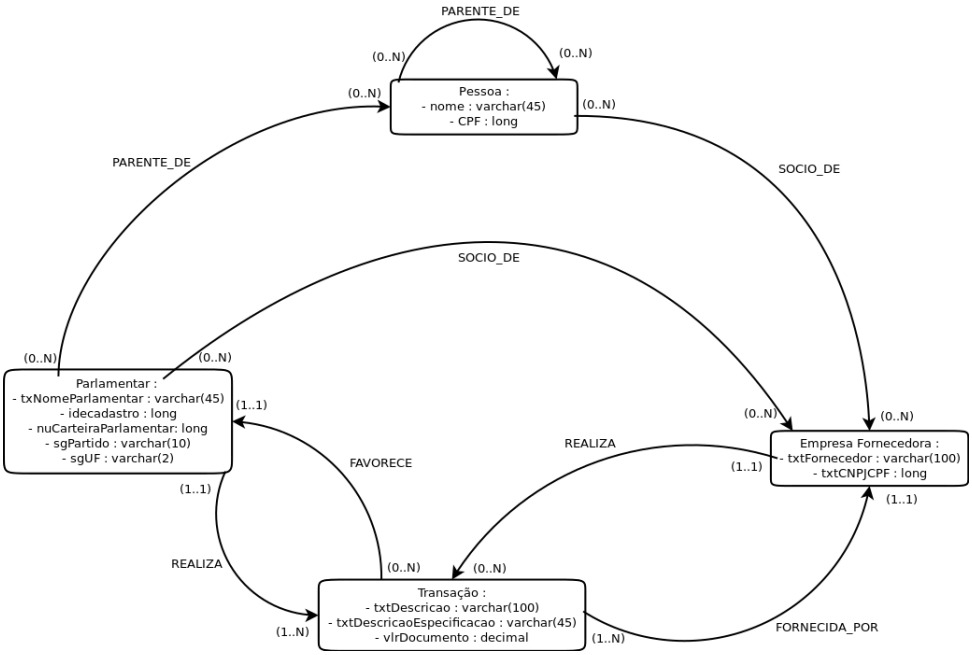

Figure 2. Modelo de dados seguindo o formato GRAPHED

Como podemos observar na Figura 2, foram propostas quatro classes que representam as instâncias dos vértices no banco de dados, que são: Parlamentar, Transação, Empresa fornecedora e Pessoa. As setas que saem de uma classe a outra, representa um

relacionamento entre essas classes, que no grafo será representado por uma aresta. Dessa forma, uma instância de um parlamentar, realiza transações, que por sua vez é fornecida por uma empresa fornecedora. Esse caminho descrito, representa as transações da CEAP, fornecida pela Câmara dos Deputados.

De forma análoga, uma empresa fornecedora realiza transações, que por sua vez favorece um certo parlamentar. Esse caminho representa os dados das doações das empresas para os deputados nas eleições de 2014, fornecido pelo TSE.

Os demais relacionamentos, como ”socio-de” e ”parente-de” tem por objetivo identificar possíveis fraudes na CEAP, uma vez que um parlamentar não pode utilizar a verba da CEAP com serviços de uma empresa que é sócio. Esses dois caminhos, se apresentaram como o maior desafio para o desenvolvimento do trabalho, uma vez que dados de parentesco dos parlamentares não são dados abertos.

3.3 ETL

Após o desenvolvimento do modelo de dados, foi desenvolvido um ETL utilizando a linguagem java, para extrair, transformar e carregar os dados para o OrientDB. O OrientDB foi desenvolvido na linguagem java, e executa portanto na JVM. Por esse motivo, o ETL foi feito em java, uma vez que, o OrientDB possui uma boa interface com essa linguagem. O ETL pode ser acessado por meio desse link no github[3].

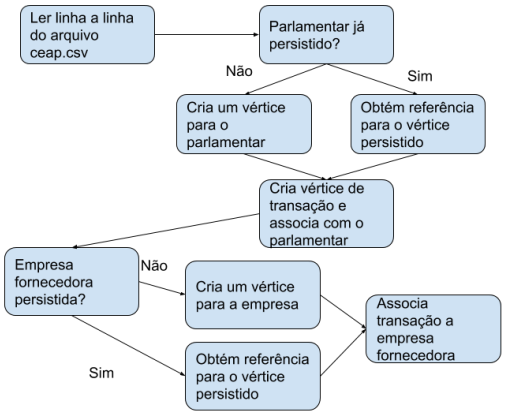

Figure 3. Fluxo de extração, transformação e carregamento para o OrientDB

O fluxo do ETL é exemplificado na Figura 3. Dessa forma o fluxo se organiza em ler linha a linha do csv fornecido pela Câmara dos Deputados, e criar os vértices e arestas com base em certas condições. Por exemplo, não se pode ter mais de um vértice representando um parlamentar, mas no arquivo fornecido existem diversas transações para um mesmo deputado.

O ponto chave desse processo são as associações entre os parlamentares com as transações e as empresas que forneceram o serviço. Criando esse vínculo, é possível por meio de consultas de casamento de padrões identificar possíveis fraudes ou padrões de transação e doação envolvendo um mesmo parlamentar e uma mesma empresa. A estrutura de um grafo fornece naturalmente os relacionamentos de forma simples e intuitiva,

[3]https://github.com/gabrielmm1234/CEAP-ETL

sendo uma ótima alternativa a estrutura relacional, que é muito comum atualmente. Além disso, a visualização da informação se torna mais clara e objetiva o que é uma ótima característica ao se tratar de transparência em dados governamentais.

O arquivo de dados da CEAP no ano de 2017 possui um total de 209496 transações, o carregamento total dessas transações levou cerca de 16 horas em um notebook com Ubuntu 16-04 LTS, Intel Core i5-5200U CPU 2.20GHz * 4 e 6 Gb de memória RAM. Esse tempo poderia ser otimizado, pois da forma que o ETL foi feito sempre é feito uma busca no banco de dados para saber se um parlamentar ou empresa já estão persistidos. Como o tempo de carga não é um objetivo prioritário no escopo do projeto essa melhora no tempo de carregamento não foi feita.

Já o arquivo das doações passou por um processo de filtragem, para obter somente os dados referentes aos deputados federais dos estados de Minas Gerais e Distrito Federal. Esse filtro para obter deputados federais dos dois estados foi feito para validar mais rapidamente a arquitetura e proposta de solução, no futuro demais estados serão adicionados no banco de dados. O arquivo já filtrado possui um total de 19302 doações de empresas a candidatura de diversos deputados. Claramente, somente alguns desses deputados foram de fato eleitos, e portanto deputados que não se encontravam na base da CEAP foram ignorados. O tempo total de processamento desse arquivo foi de 5 horas, podendo ser melhorado nos mesmos princípios dos dados da CEAP.

3.4 Resultados das consultas

Finalizado a obtenção e carregamento dos dados no OrientDB, foram feitas consultas em busca dos relacionamentos entre as empresas doadoras e os deputados de Minas Geras e Distrito Federal que utilizaram a CEAP com essas mesmas empresas. O OrientDB possui uma linguagem de consulta baseada na linguagem SQL, diferentemente de outros SGBD orientado a grafos como o Neo4J, que possui uma linguagem própria conhecida como Cypher.

A consulta feita para obter os resultados esperados, é baseada no conceito de casamento de padrão, muito comum em linguagens funcionais. Dessa forma, para obter um certo padrão dentro do grafo o OrientDB fornece a função MATCH, como mostra a Consulta 1.

```
MATCH
  {class:Parlamentar, as:p} -RealizaTransacao->
        {class:Transacao, as:t}
      -FornecidaPor-> {class:EmpresaFornecedora, as:e},
  {as:e} -RealizaTransacao-> {class:Transacao, as:t2}
           -FornecidaPara-> {as:p}
RETURN $elements
```

Listing 1. Consulta de relacionamento de doações entre deputados e empresas

A Consulta 1 busca um certo padrão dentro do banco de dados. Esse padrão é definido por parlamentares chamados de ”p”, que realizam transações da CEAP ”t”, fornecidas por uma certa empresa ”e”. Além disso essa mesma empresa ”e”, realiza uma transação de doação ”t2”, que é fornecida para o mesmo parlamentar ”p” definido no

início da sentença. O resultado dessa consulta nos deputados de Minas Gerais e Distrito federal é apresentado na Figura 4.

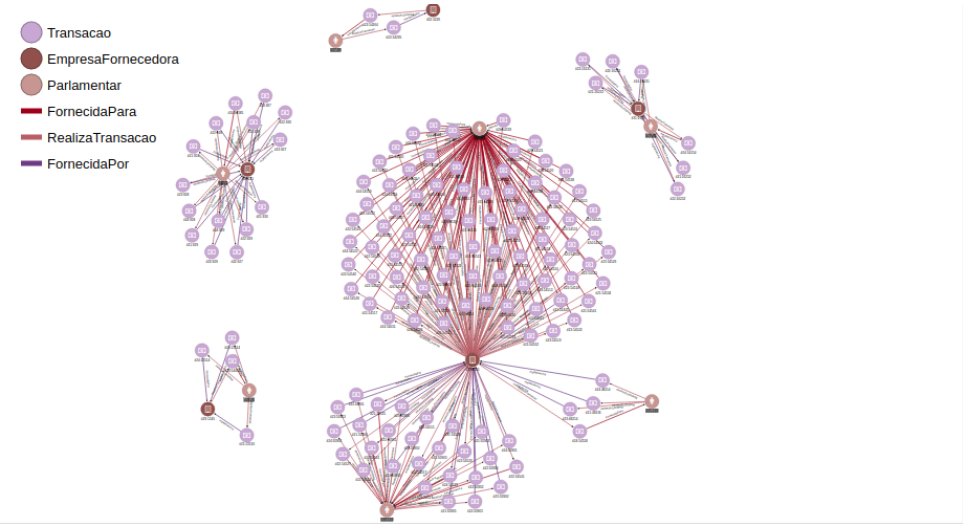

Figure 4. Resultado do cruzamento entre dados da CEAP e do TSE

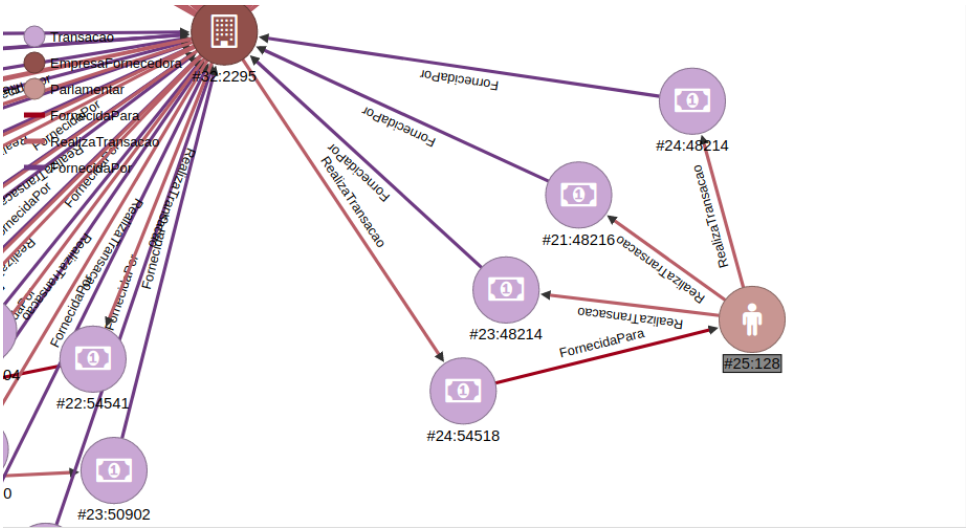

Figure 5. Padrão de relacionamento entre dados da CEAP e do TSE

O padrão definido foi encontrado para um total de sete parlamentares todos em Minas Gerais. A Figura 5 mostra em mais detalhes como o padrão é definido.A partir da orientação das setas, é possível perceber que o parlamentar representado pelo vértice marrom claro e com o ícone de uma pessoa, realizou três transações, representadas pelo vértice roxo claro, com uma certa empresa representada pelo vértice marrom escuro. Sendo que essa empresa fez uma doação para esse parlamentar. No total, foram 44 transações registradas na CEAP, e 96 doações registradas pelo TSE que estão seguindo esse padrão.

Foi feita uma consulta para obter os padrões que podem ser considerados fraudes na CEAP, e testada com dados fictícios. No caso a consulta busca por deputados que usaram a CEAP com empresas das quais são sócios, como apresentada na consulta 2.

```
MATCH
  {class:Parlamentar, as:p} -RealizaTransacao->
        {class:Transacao, as:t}
        -FornecidaPor-> {class:EmpresaFornecedora, as:e},
  {as:p} -Socio_De-> {as:e}
RETURN $elements
```

Listing 2. Consulta de relacionamento de uso da CEAP entre deputados e empresas nas quais o deputado é sócio

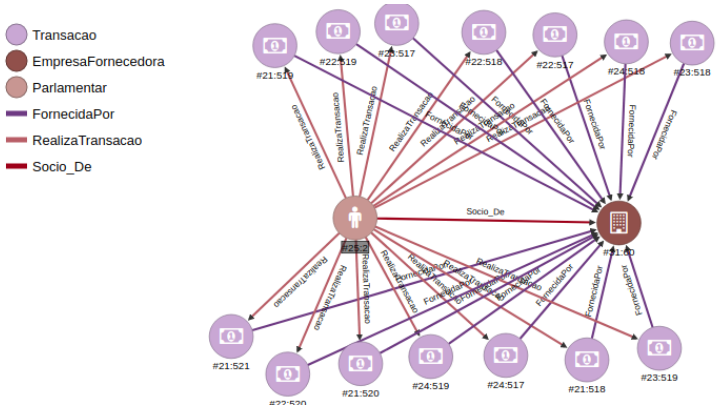

Figure 6. Padrão de uma transação fraudulenta com dados fictícios

Como mostra a figura 6, a consulta consegue localizar um padrão de transações efetuadas entre um deputado e uma empresa, na qual o deputado faz parte do quadro de sócios. A forma com que o grafo é apresentado visualmente, pode facilitar bastante o entendimento dos relacionamentos, de forma que fica bastante claro que um deputado, representado pelo círculo com o ícone de uma pessoa, é sócio de uma empresa que prestou serviços com o dinheiro da CEAP.

4 Conclusões e trabalhos futuros

Esse trabalho apresentou, portanto, o uso de banco de dados em grafo para detectar indícios de fraudes na cota para exercício parlamentar, bem como relacionamentos entre esses dados e os dados de doações para as campanhas eleitorais em 2014. Além disso, demonstra o desafio de se trabalhar com dados abertos no Brasil, uma vez que, os dados nem sempre são padronizados. Também mostra o uso dos bancos de dados NoSQL com dados abertos, e em aplicações que tem por objetivo apresentar informações relevantes a sociedade de forma objetiva e transparente.

Os resultados mostram como é intuitivo observar resultados olhando para a estrutura de um grafo, em vez de analisar uma tabela. Isso coloca os bancos de dados em grafos como ótima opção para aplicações que tem por objetivo a transparência e interpretação de dados complexos e bastante relacionados.

Para trabalhos futuros está incluso a implementação de um sistema colaborativo junto com a detecção de fraudes na CEAP. Este sistema irá utilizar o banco de dados desenvolvido neste trabalho, de forma a possibilitar que a população visualize diversas informações a respeito da CEAP, e possa contribuir com informações importantes nas

consultas que detectam as fraudes, como por exemplo, dados de parentesco dos deputados, quadro societário das empresas fornecedoras e etc. Dessa forma, quanto mais informações forem fornecidas maior a chance de se encontrar uma transação fraudulenta.

Além disso, temos o uso de técnicas de aprendizagem de máquina nos dados da CEAP, que traz um problema interessante de executar essas técnicas em uma arquitetura baseada em um SGBD orientado a grafos, para encontrar padrões e criar modelos preditivos. Outro ponto importante, é estudar as formas de visualização de grafos na web e o impacto do uso de grafos como ferramenta de visualização para dados abertos, que estão relacionados com o sistema colaborativo mencionado acima.

References

[Abramo 2000] Abramo, C. W. (2000). Relações entre índices de percepção de corrupção e outros indicadores em onze países da américa latina. *SPECK, Bruno W. et al. Os custos da corrupção. Cadernos Adenauer*, (10):47–62.

[Barmpis and Kolovos 2014] Barmpis, K. and Kolovos, D. S. (2014). Evaluation of contemporary graph databases for efficient persistence of large-scale models. *Journal of Object Technology*, 13(3):3–1.

[Bondy et al. 1976] Bondy, J. A., Murty, U. S. R., et al. (1976). *Graph theory with applications*, volume 290. Citeseer.

[Diirr and Cappelli] Diirr, B. and Cappelli, C. Combate à corrupção através de relacionamentos interorganizacionais transparentes.

[Filgueiras 2009] Filgueiras, F. (2009). A tolerância à corrupção no brasil: uma antinomia entre normas morais e prática social. *Opinião Pública*, 15(2):386–421.

[Gelbmann 2014] Gelbmann, M. (2014). Graph dbmss are gaining in popularity faster than any other database category. `https://db-engines.com/en/blog_post/26`. Acessado em janeiro de 2018.

[Labute and Dombroski 2014] Labute, M. and Dombroski, M. (2014). Review of graph databases for big data dynamic entity scoring. Technical report, Lawrence Livermore National Laboratory (LLNL), Livermore, CA.

[Nayak et al. 2013] Nayak, A., Poriya, A., and Poojary, D. (2013). Type of nosql databases and its comparison with relational databases. *International Journal of Applied Information Systems*, 5(4):16–19.

[Van Erven et al. 2018] Van Erven, G., Silva, W., Carvalho, R., and Holanda, M. (2018). Graphed: A graph description diagram for graph databases. In Rocha, Á., Adeli, H., Reis, L. P., and Costanzo, S., editors, *Trends and Advances in Information Systems and Technologies*, pages 1141–1151, Cham. Springer International Publishing.