Modelagem em Grafos a partir de Bancos de Dados Relacionais

Silas P. Lima Filho
Instituto Militar de Engenharia
Rio de Janeiro, RJ, Brasil
Email: silasplfilho@outlook.com

Maria C. Cavalcanti
Instituto Militar de Engenharia
Rio de Janeiro, RJ, Brasil
Email: yoko@ime.eb.br

Claudia M. Justel
Instituto Militar de Engenharia
Rio de Janeiro, RJ, Brasil
Email: cjustel@ime.eb.br

***Resumo*—A importância de trazer dados do modelo relacional para outros modelos e tecnologias tem sido amplamente debatidos, como por exemplo a publicação de dados como grafos. Este modelo permite executar análises topológicas, tal como ocorre nas análises de redes sociais, predição de ligações e sistemas de recomendação. Existem iniciativas para mapear de um banco de dados relacional para a representação em grafo. No entanto, eles não consideram as diferentes maneiras de gerar esses grafos, especialmente quando o objetivo é realizar análises topológicas. Este trabalho propõe heurísticas para facilitar a sistematização do mapeamento de dados do modelo relacional para a representação em grafos. A principal contribuição é que a escolha do modelo do grafo deve considerar o tipo de análise topológica que será realizada pelo usuário. Experimentos são apresentados e mostram resultados interessantes, incluindo heurísticas para apoiar o usuário na escolha da modelagem do grafo.**

1. Introdução

Uma das principais razões para converter dados armazenados em bancos de dados relacionais em uma representação em grafo, é para facilitar a tomada de decisão. Por décadas, o modelo de grafos tem sido amplamente explorado para fins de análises topológicas. Existem vários algoritmos e aplicações na literatura. Normalmente, esses algoritmos são utilizados no contexto de redes sociais, sistemas de recomendação, predição de ligações e outros.

Com o crescimento da disponibilidade de grafos na Web de Dados, e com o surgimento de novas tecnologias, o uso de grafos e parte de sua teoria tem sido aplicada em dados disponíveis na Web, como na iniciativa *Linked Data*. Nessa direção, algumas alternativas para mapear dados relacionais para grafos RDF surgiram, como o D2R [1]. No entanto, essa abordagem foca em um mapeamento sintático e não nas questões de modelagem de grafos. Por outro lado, existem trabalhos que analisam a modelagem de grafos, como por exemplo [2], [3] e [4]. No entanto, os autores utilizam uma abordagem específica que não leva em consideração alternativas de modelagem e as consequências da modelagem escolhida em alguma análise topológica.

O objetivo desse trabalho é mostrar que modelagens diferentes podem levar a conclusões diferentes. Em outras palavras, duas modelagens derivadas do mesmo conjunto de dados, podem identificar diferentes conjuntos de nós como mais centrais. A ideia é apoiar o usuário na consideração de modelagens alternativas, identificando, a partir de um esquema conceitual do banco de dados, o que precisa ser explicitado como nó do grafo, de modo a permitir que as análises topológicas tragam resultados úteis. A principal contribuição deste trabalho é a identificação de heurísticas que levam em consideração as análises topológicas intencionadas. Elas foram identificadas com base nos resultados obtidos de experimentos de análise topológica (e.g. medidas de centralidade) sobre diferentes modelagens.

Este artigo se divide da seguinte maneira. Na seção 2 estão apresentados os elementos necessários para o entendimento do trabalho. Na seção 3 estão relacionados os trabalhos que abordam o mesmo problema abordado. O método utilizado para mapear um banco relacional para um grafo é apresentado na seção 4, tal como o modo como as heurísticas foram construídas também é apresentado nessa seção. Em sequência, os experimentos iniciais, com conjuntos de dados menores são detalhados na seção 5. Um experimentos maior com o intuito de confirmar o comportamento dos experimentos anteriores é feito na seção 5.3. Finalizando na seção 6 com a conclusão do trabalho realizado com pontos destacados, possíveis de trabalhos futuros.

2. Fundamentos Básicos

Nessa seção, apresentamos parte do fundamento teórico necessário para o desenvolvimento do trabalho.

2.1. Análise de Redes Sociais

As redes sociais têm sido amplamente usadas nos últimos tempos, devido a sua aplicação na solução de diversos problemas que envolvem relacionamentos entre pessoas ou objetos. Existem diferentes maneiras de analisar uma rede social. O foco deste trabalho é fazer uma análise topológica de uma rede. Ou seja, a rede será modelada como um grafo, e aplicaremos algoritmos para determinar informações sobre grau de nós, caminhos e distâncias entre pares de nós. Uma das possíveis análises topológicas consiste em determinar a influencia dos nós na rede levando

em consideração a quantidade de conexões entre eles. Essas conexões (relações) permitem determinar quais nós são mais ou menos relevantes, ou podem transmitir informação que trafega pela rede. Diferentes medidas de centralidade foram definidas na literatura para permitir essas análises. Apresentamos a seguir as medidas que serão utilizadas neste trabalho.

Seja $G = (V(G), E(G))$ um grafo que representa uma rede. A centralidade de **Grau** é simplesmente a medida definida como a quantidade de arestas conectadas a um nó $x \in V(G)$, formalmente, $C_d(x) = |\Gamma(x)|$, onde $\Gamma(x)$ é o subconjunto de vértices em $V(G)$ adjacentes a x. A medida **Proximidade** leva em consideração a distância de um determinado nó para todos os outros nós do grafo e é definida como a soma dos inversos das distâncias entre x e todos os nós do grafo. Formalmente $C_c(x) = \sum_{s \in V(G), s \neq x} \frac{1}{d(x,s)}$, onde $d(x, s)$ é a distância (comprimento do menor caminho) entre os nós x, s em $V(G)$. A centralidade de **Intermediação** mede a quantidade de vezes que um nó x estará presente no menor caminho entre outros dois nós do grafo, s, t. Nós com valor de intermediação alto são mais propensos a controlar o fluxo de informação no grafo. Formalmente, $C_b(x) = \sum_{s \neq t \neq x \in V(G)} \frac{n^x_{s,t}}{g_{s,t}}$, onde $n^x_{s,t}$ é o número de caminhos mínimos entre os nós s e t que contém o nó x; e $g_{s,t}$ é o número total de caminhos mínimos entre os nós s e t, $\forall s \neq t \in V(G)$. As medidas de centralidade Hub e Autoridade são definidas para redes direcionadas. Newman [5] afirma que nós com alta centralidade são aqueles que são apontados por outros nós com centralidade alta. Porém também é possível ter um nó com alta centralidade quando ele aponta para nós de alta centralidade. Dessa forma podem ser definidos dois tipos de nós centrais numa rede direcionada, **Authoridade** e **Hub**. Esses conceitos foram criados para identificar páginas relevantes na Web, baseados na ideia de reconhecer páginas com muitas conexões a outras páginas que por sua vez altamente conectadas.

2.2. Modelagem de Bancos Relacionais e de Grafos

Heuser [6] afirma que um modelo de dados deve ser expressivo o suficiente para criar esquemas de dados, isto é, de modo a refletir objetos do mundo real. Modelar um esquema de banco de dados é uma tarefa que normalmente abrange duas etapas com diferentes níveis de abstração: conceitual e lógico. Estas duas etapas são necessárias em razão da complexidade da realidade que o usuário encarregado da modelagem pretende representar. A idéia principal é conduzir o usuário desde a realidade até alguma estrutura de dados lógica que pode representar objetos reais. O Modelo ER [7] é frequentemente usado para modelar esquemas conceituais, onde objetos e relacionamentos do mundo real são representados como entidades (classes de objetos) e seus tipos de relacionamento (veja a Figura 1). Em um segundo momento estas entidades e relacionamentos são mapeados para um esquema lógico. O Modelo Relacional [8] é amplamente usado para criar esquemas lógicos, onde tabelas são definidas como estruturas que irão efetivamente armazenar os dados.

Diferente da modelagem de um banco de dados, cujo alvo é o armazenamento e gerenciamento de dados, a modelagem em grafos é direcionada para a análise topológica dos dados, como no caso de análise usando medidas de centralidade. A necessidade de atender ambos os objetivos, o armazenamento/gerenciamento e a análise topológica, permitiu a criação dos Sistemas de Gerenciamento de Banco de Dados Orientado a Grafos (SGBDG). Esses sistemas usam uma estrutura de um grafo para armazenar os dados. Neo4J[1] é um dos GDBMS mais usados na atualidade.

Em [9] Rodriguez analisa diferentes estruturas de grafos, algumas das quais permitem representar diferentes características para os nós e arestas, tais como rótulos, atributos, peso, etc. Nesse artigo, é apresentada uma classificação hierárquica, onde as estruturas de grafos são organizadas de acordo com a sua expressividade (número de características permitidas). O **grafo de propriedades** (*property graph*), é a estrutura mais usada por algumas ferramentas de manipulação (e.g. Neo4j), devido a sua alta expressividade.

Grafos podem ser modelados a partir de alguns elementos de um banco de dados. Para atingir o objetivo de mapear os elementos de um banco de dados relacional para um grafo, é necessário utilizar os dois esquemas, conceitual e lógico. O usuário pode usar esses esquemas para identificar quais itens serão representados como vértices e que tipo de referências serão arestas no grafo. Porem, essa não é uma tarefa fácil, especialmente, quando há diversas análises que podem ser realizadas. No entanto, até aonde foi possível investigar, não achamos métodos ou guias que ajudem o usuário em tal tarefa, levando em conta a análise desejada sobre o grafo.

3. Trabalhos Relacionados

Na literatura existem métodos diferentes com o objetivo de modelar um grafo a partir de um modelo de banco de dados relacional. Por exemplo, em [2] é tratado o problema de modelagem de um grafo procurando pela melhora da performance de uma consulta num banco de dados em grafos. Outro artigo, [3], procura evitar perda semântica durante a modelagem. Já em [4] o objetivo é evitar redundância nos dados.

A idia principal do artigo de Wardani e Küng, [3], é construir um grafo o mais próximo possível ao esquema conceitual do banco de dados relacional, evitando perdas semânticas. Os autores usam os dois esquemas, conceitual e relacional, para criar regras de mapeamento específicas, como o uso de atributos de chave estrangeira para mapear $a(n)$ relacionamentos/arestas entre pares de nós. A partir dessas regras, o grafo pode ser construído.

De maneira semelhante, o artigo de De Virgilio et al., [10], começa com uma análise do esquema relacional. E num outro artigo, [2], os mesmos autores destacam a necessidade de fazer uma análise cuidadosa a partir do esquema

1. https://neo4j.com/developer/

conceitual ER, de modo a resolver possíveis conflitos ao agregar entidades e relacionamentos. Para tal fim, propõe primeiramente o resgate do esquema conceitual ER a partir do esquema relacional. Com base nessa representação é concebido um grafo *"template"* associado ao esquema lógico de dados a serem utilizados. Nesse esquema, as entidades e relacionamentos são convenientemente agrupados, respeitando as restrições de integridade referencial definidas no esquema relacional, e algumas regras definidas pelo autor. A ideia é agrupar entidades cujas instâncias irão aparecer juntas em algum resultado de consulta, otimizando dessa forma o processamento dessas consultas.

Enquanto os métodos previamente descritos propõem a criação de um grafo de propriedades, Bordoloi et al., [4], apresentam um método de construção de um hipergrafo que parte de um esquema relacional. Inicialmente, são construídos grafos estrela e grafos de dependência, mostrando as relações de dependência entre os atributos para cada tabela. Depois, esses grafos são unidos formando um único hipergrafo. Semelhante ao grafo chamado *"template"* no artigo de De Virgilio, [2], este hipergrafo representa o esquema do banco, onde os nós são os atributos das tabelas, e as arestas representam as dependências entre os atributos. Com base nesse hipergrafo, poderia ser gerado o hipergrafo dos dados, onde cada valor de atributo nas tuplas das relações passa a ser um nó, e as relações de dependência também são instanciadas. No entanto, este é um grafo complexo, com muitos nós. Para reduzir este grafo e evitar redundância de nós, o método proposto inclui uma análise de domínios comuns entre os atributos. Então, outro hipergrafo esquema é construído levando em conta aquela análise, no qual atributos do mesmo domínio, que apareçam em tabelas diferentes, são representados de maneira única. Finalmente, um hipergrafo de dados, no qual um nó representa um valor de um domínio específico, é construído baseado no hipergrafo esquema. Embora nesta abordagem todos os valores atributos podem ser analisados, o hipergrafo não é fácil de tratar, pois muitas das ferramentas e algoritmos não admitem este tipo de estrutura como entrada.

Um trabalho recente de Xirogiannopoulos et al., [11], apresenta uma ferramenta de extração de dados de um banco de dados relacional para um modelo em grafo, denominada *GraphGen*, que permite ao usuário realizar análise do grafo, ou executar algoritmos usando o grafo obtido como entrada. Essa ferramenta utiliza a linguagem DSL, que é baseada em *Datalog*, para realizar as extrações do banco relacional. Este trabalho é o único dos achados que discute a relevância de obter diferentes modelos de grafos a partir do mesmo conjunto de dados. Porém, não conduz o usuário para fazer a escolha desses modelos.

No entanto, mesmo que o usuário consiga fazer análises sobre o grafo extraído (modelagem escolhida), não há um estudo sobre o impacto dessa escolha, nos resultados das análises. Assim, embora o usuário tenha a liberdade de escolher a modelagem do grafo, sente-se a necessidade de garantir que ele explore as alternativas, identificando entre as mesmas, aquela que será mais útil para suas análises.

4. Preparação para os Experimentos

No sentido de guiar o usuário na escolha da modelagem em grafo entre as alternativas, um conjunto de experimentos foi realizado, e estes estão relatados na próxima seção. Para realizá-los, foi necessário tomar como ponto de partida um esquema conceitual (**EC**). Usualmente, um esquema conceitual tem várias entidades e relacionamentos, mas para iniciar nosso estudo foram escolhidas duas situações de modelagem comuns (recortes), que aparecem com frequência nos projetos de banco de dados. O primeiro recorte escolhido *(a)* consiste de **um par de entidades** (E_1, E_2) com **n relacionamentos**. O segundo recorte escolhido *(b)*, envolve **um par de entidades** (E_1, E_2) e apenas **um relacionamento binário n:m R**, ao qual está associado um *atributo multivalorado A*.

As Figuras 1 e 2 mostram estes recortes utilizando o modelo ER. Neste trabalho, apenas estes dois recortes foram analisados. Porém, vale mencionar, que um esquema pode ser visto como um conjunto de recortes, que podem ser analisados, um de cada vez, de modo incremental.

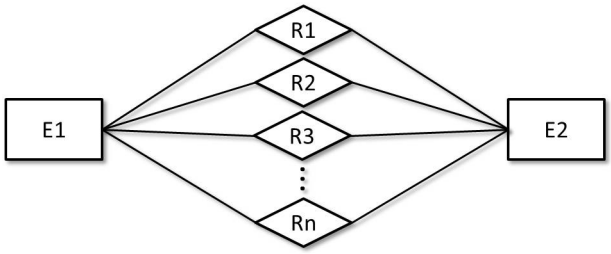

Figura 1. Situação típica em esquemas conceituais: duas entidades e vários relacionamentos.

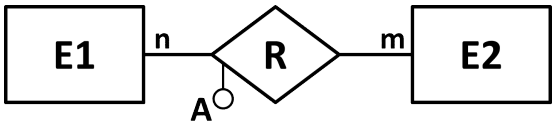

Figura 2. Situação típica em esquemas conceituais: duas entidades e um relacionamento n:m.

Os experimentos deram-se da seguinte forma: escolheu-se um banco de dados, e a partir de um recorte de seu esquema conceitual (extraído a partir de seu esquema lógico) foram geradas diferentes alternativas de modelagem em grafo. Para cada grafo, foi aplicado um conjunto de análises topológicas. O objetivo desses experimentos foi avaliar o impacto que cada alternativa de modelagem teve nos resultados de análises realizadas sobre elas. E, a partir desta avaliação, foi possível identificar heurísticas genéricas que visam ajudar neste processo. Em outras palavras, essas heurísticas poderão ser usadas para apoiar os usuários na escolha da modelagem mais útil.

A Figura 3 exibe um fluxo de construção de uma heurística. A partir de um conjunto de dados (BD) e de um esquema conceitual (EC), são geradas diferentes alternativas

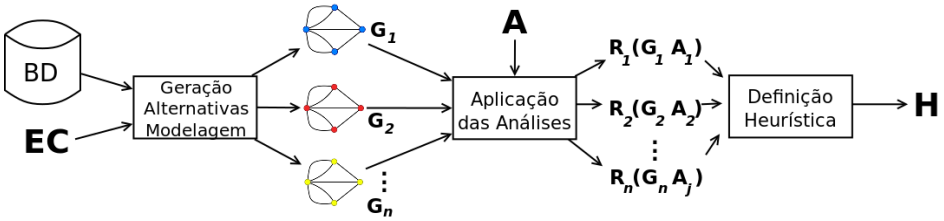

Figura 3. Fluxo para definição de uma heurística em um experimento

de modelagem de grafos $(G_1, G_2, ..., G_n)$, e para cada uma dessas alternativas, são aplicadas análises pré-selecionadas $(A_1, A_2, ..., A_n)$. A partir dos resultados de cada grafo, $(R_i(G_i, A_i), 1 \leq i \leq n)$, foi possível compará-los e definir uma heurística (H).

5. Experimentos

As próximas subseções apresentam os experimentos realizados e seus resultados, bem como as heurísticas identificadas. Os bancos de dados escolhidos para os primeiros experimentos foram intencionalmente de pequeno porte, para explorar como as diferentes alternativas de modelagem de dados em grafos podem impactar nas análises que se objetiva fazer. Embora existam diferentes tipos de grafos, para este trabalho foi utilizado o grafo de propriedades, apresentado na seção 2.2. Os motivos da escolha são: permitir uma maior liberdade de modelagem e o fato desse tipo de grafo ser usado largamente na literatura.

Os experimentos foram realizados com grau de dificuldade crescente, aumentando também a quantidade de dados usados. As análises topológicas escolhidas para os experimentos foram medidas de centralidade, que são comumente usadas para análise de redes sociais.

5.1. Experimento Zachary

No experimento Zachary, utilizou-se uma base de dados sobre os relacionamentos dos membros de uma academia de karatê criado a partir da rede de Zachary [12]. Nesse conjunto, os membros de um clube de karatê podem relacionar-se de duas formas, pelo relacionamento *"Amigo de"* ou então pelo relacionamento *"Lutou com"*, este último visando representar resultados de lutas em campeonatos e torneios.

A Figura 4 representa o esquema conceitual do banco relacional usado nesse experimento. O esquema é similar ao recorte (a), onde $E_1 == E_2 == Pessoa$ e os auto-relacionamentos *"Amigo de"* e *"Lutou com"* do esquema, representam os dois tipos de relacionamento que podem existir entre as pessoas.

Esse experimento foi dividido em duas etapas para atender às duas alternativas de modelagem a serem investigadas. Em ambas foram analisadas algumas medidas de centralidade, sendo elas intermediação, proximidade e grau. Na primeira etapa, o grafo analisado (Modelagem 1), que é apresentado na Figura 5, foi gerado com nós representando

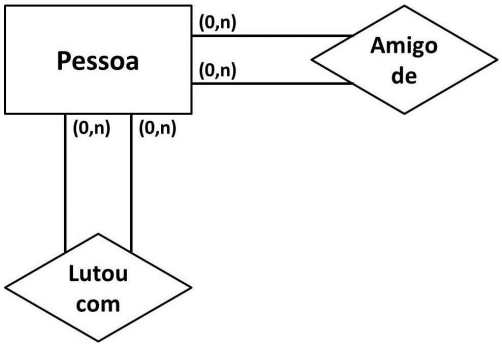

Figura 4. Esquema conceitual do Banco de Dados para o Experimento Zachary.

pessoas, e arestas representando ambos os tipos de relacionamento. Na segunda etapa, o grafo analisado (Modelagem 2) foi similar ao primeiro, mas neste não foram consideradas as arestas que representam o relacionamento *"Lutou com"*, mantendo assim a rede de amizade original do conjunto de Zachary, como mostra a Figura 6. O resultado das análises realizadas com as duas modelagens pode ser conferido na Tabela 1.

Para o armazenamento dos dados, utilizou-se o software Neo4j. Enquanto que para o cálculo das medidas utilizou-se a ferramenta de manipulação e visualização de grafos, Gephi.

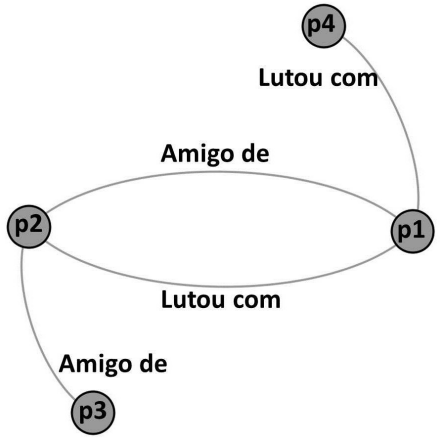

Figura 5. Modelagem 1 para o Experimento Zachary.

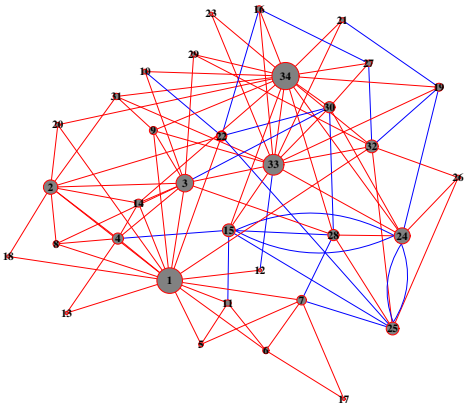

Figura 6. Modelagem 2 para o Experimento Zachary.

	Modelagem 1			Modelagem 2		
Nó	**Intermediação**	**Proximidade**	**Grau**	**Intermediação**	**Proximidade**	**Grau**
1	152,19	1,6	16	231,0	1,758	16
34	102,99	1,66	17	160,5	1,81	17
4	12,5	2	7	6,2	2,15	6
22	24	1,9	6	0	2,6	2
24	14,6	2,0	8	9,3	2,5	5
30	9,7	1,9	7	1,5	2,6	4

Tabela 1. RESULTADOS DAS MEDIDAS DE CENTRALIDADE PARA AS DUAS MODELAGENS DO EXPERIMENTO ZACHARY.

5.1.1. Análise dos Resultados. Observando os resultados na Tabela 1 podemos concluir que o experimento demonstra a influência da modelagem escolhida nas medidas de centralidade obtidas. Dependendo da modelagem escolhida, informações importantes para a análise do usuário devem ser consideradas ou não. No caso da Modelagem 2, onde são analisados apenas os relacionamentos de amizade, os nós 1 e 34 são os mais centrais segundo a medida de centralidade intermediação. Na Modelagem 1, por mais que os nós mais centrais continuem os mesmos, houve uma mudança significativa nos valores de centralidade dos nós 4, 22, 24 e 30. Esta mudança pode-se interpretar da seguinte forma, tais nós representam pessoas que participaram de vários torneios e lutas, como é o caso do nó 22, cujo valor do grau e de intermediação teve a maior alteração. Assim, caso o modelador deseje encontrar os lutadores de maior visibilidade social no grafo, deve levantar todas as relações envolvendo os lutadores. Se achar que a relação de amizade é a mais útil para identificar os lutadores, então esta deve ser considerada. Mas, se a visibilidade social pode ser alcançada também com as relações de luta, esta deve ser considerada além da relação de amizade. Assim, ambas as relações (amizade e luta) devem ser representadas no grafo.

De modo mais genérico, esta situação de modelagem pode envolver duas entidades distintas, que possuam mais de um tipo de relacionamento entre elas. Em casos onde existam muito mais que dois tipos de relacionamentos entre duas entidades, a consequência gerada nos valores das medidas de centralidade pode ser muito mais significativa. Em tais casos, uma análise cuidadosa por parte do usuário sobre quais relacionamentos são semanticamente relevantes, deve ser feita.

À medida que mais tipos de relações são considerados, mais arestas surgirão no grafo, e naturalmente as medidas de centralidade serão impactadas. Assim, se o objetivo é encontrar os nós mais centrais ou de maior intermediação, por exemplo, é preciso analisar para o contexto dos dados em questão, quais os tipos de relacionamento que devem ser levados em consideração. Em geral, há uma diferença semântica entre os dois tipos de relacionamento, e é o que justifica o fato de serem representados de modo distinto no banco de dados. Mas, se para fins da análise da rede, ambos podem ser considerados semanticamente relevantes, então ambos devem ser representados no grafo.

Embora no banco de dados usado no experimento, apenas uma entidade tenha sido considerada, de modo mais genérico, a mesma observação deve ser feita em uma situação de modelagem onde banco de dados envolve duas entidades distintas, que possuam mais de um tipo de relacionamento entre elas.

Mesmo usando algumas medidas de centralidade para descobrir quais são os nós mais centrais, a descoberta de tais nós não chegou a ser um fator de influência direta para a obtenção da primeira heurística. O valor de cada medida de centralidade de cada nó foi comparado para as duas modelagens. E desse estado então, foi compreendido que a escolha de relacionamentos pode interferir na determinação dos nós mais centrais do grafo. A conclusão mais relevante neste experimento foi o entendimento da diferença entre relacionamentos e sua semântica, pois dependendo da forma de representação, afetará o resultado da análise das medidas de centralidade utilizadas.

5.1.2. Heurística 1. A partir da conclusão obtida no experimento Zachary, foi obtida uma heurística que será apresentada a seguir. Dadas duas entidades E_1 e E_2 de um esquema conceitual EC, não necessariamente distintas, que se relacionam por relacionamentos distintos $R_1, R_2, ..., R_n$. Se a análise a ser realizada tem como foco um relacionamento R_i a ser modelado no grafo, então para todo $j \neq i$ tal que R_j tenha a mesma relevância semântica[2] que R_i, R_j também deve ser modelado no grafo. Ou mais formalmente, como se segue.

Sejam:

- $G = (V, A)$; grafo
- $EC = (E, R)$; esquema conceitual
- $E = \{E_1, E_2, ..., E_p\}$; conjunto de entidades do esquema EC
- $R = \{R_1, R_2, ..., R_q\}$; conjunto de relacionamentos do esquema EC
- $E_i = \{e_i^1, e_i^2, ...\}$; conjunto de instâncias da entidade E_i
- $R_k = \{r_k^1, r_k^2, ...\}$; conjunto de instâncias do relacionamento R_k
- $r_k^z = (e_i^x, e_j^y)$; um instância de R conecta um par de instâncias de entidades de E

Dado que:

2. Obs: o simbolo $\approx$ foi usado com o significado de semanticamente relevante.

- $E_1, E_2 \in E; e_1^x \in E_1; e_2^y \in E_2;$
- $R_1 \in R;$
- $r_1^z \in R_1; r_1^z = (e_1^x, e_2^y);$

No mapeamento para o grafo G temos:

- $\forall e_1^x \in E_1, e_1^x \in V;$
- $\forall e_2^y \in E_2, e_2^y \in V;$
- $\forall r_1^z \in R_1, r_1^z \in A;$
- Se $\exists j \neq 1 \,|\, r_j^c = (e_1^a, e_2^b)$ e $R_j \approx R_1$ então $\forall r_j^c \in R_j; r_j^c \in A$

A Figura 7 ilustra a escolha dos relacionamentos $R_i, ...R_j$ entre as entidades E_1 e E_2 do esquema conceitual que serão utilizados para construir um grafo onde cada nó representa uma entidade, e cada aresta representa um dos relacionamentos escolhidos.

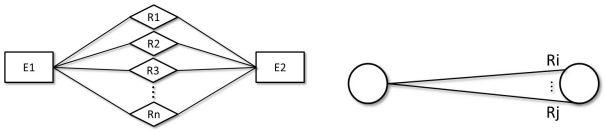

Figura 7. Representação da heurística obtida no Experimento Zachary

5.2. Experimento SMDB

Nesse experimento, utilizou-se um subconjunto de dados do banco de dados disponibilizado pela ferramenta Neo4j para introduzir a linguagem *Cypher*, responsável por realizar as consultas no banco. Nesse banco, são mantidas informações sobre dois tipos de entidades (pessoas e filmes), e os distintos relacionamentos entre essas duas entidades. Foram usados apenas os relacionamentos sobre ações de atuação e direção das pessoas dessa base. O esquema conceitual do banco relacional pode ser visto na figura 8.

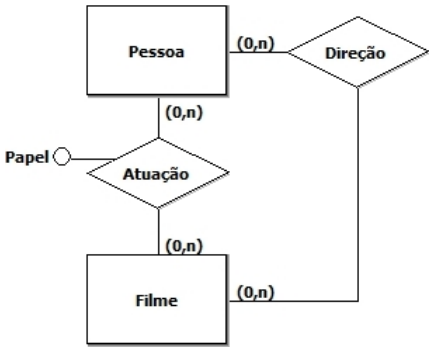

Figura 8. Esquema conceitual do Banco de Dados para o Experimento SMDB.

Para a realização desse experimento, não foram usadas as mesmas ferramentas do experimento Zachary. As medidas de centralidade foram calculadas usando Gephi [3] e os dados foram armazenados usando Neo4j. Nesse experimento, foi analisado uma situação de modelagem semelhante ao recorte *(b)*, descrito na seção 4, onde o esquema conceitual possui um par de entidades (E_1, E_2) ligados por **dois relacionamentos binários**.

Embora a situação de modelagem de BD seja semelhante à situação do primeiro experimento, neste caso, temos duas entidades distintas, além de dois relacionamentos distintos a considerar. A partir desta situação de modelagem de BD foi possível explorar três variações de modelagens em grafo, considerando a explicitação das relações como nós, e de atributos como nós.

Assim, esse experimento foi dividido em três etapas. Cada etapa considerou uma forma de migração, isto é, uma modelagem em grafo. Em todas as etapas foram analisadas as seguintes medidas topológicas para os nós do grafo obtido: intermediação e proximidade. Foram também medidos o valor de *hub* e *autoridade* de cada nó, utilizando a orientação das arestas do grafo.

Na primeira etapa, da mesma forma que no experimento Zachary, as entidades do esquema conceitual são transformadas em nós no grafo. Agora, existem dois tipos de nós, *Pessoa* e *Filme*, e os relacionamentos são transformados em arestas no grafo. Também agora existem dois tipos de relacionamento, *Atuação* e *Direção*. A Figura 9 ilustra o grafo obtido nessa etapa **(Modelagem 3)**. No caso do relacionamento *Atuação* do modelo ER, temos um atributo *Papel* associado. Essa informação é mantida no grafo como atributo de aresta.

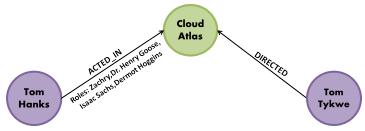

Figura 9. Esquema do grafo para a Modelagem 3

Na segunda etapa, diferente do experimento Zachary, os relacionamentos foram transformados em nós. Assim, o grafo nessa etapa possui quatro tipos distintos de nós. Nessa modelagem os relacionamentos não possuem atributos ou rótulos, todos eles estão no mesmo nível. Existem arestas entre os nós *Pessoa* e *Atuação*, *Atuação* e *Filme*, *Filme* e *Direção* e entre *Direção* e *Filme*. A Figura 10 representa o esquema do grafo (**Modelagem 4**).

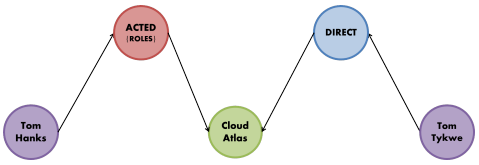

Figura 10. Esquema do grafo para a Modelagem 4

Na terceira etapa, a modelagem mantém os mesmos tipos de nós e arestas que a **modelagem 4**, mas foi inserido um novo tipo de nó que representa cada tipo de atuação que um mesmo ator teve no filme, *Papel*, e novos tipos de aresta entre nós *Ator* e *Papel*, e entre os nós *Papel* e *Filme*. Isto é, no caso dos nós do tipo *Atuou_em* em que um ator fez

3. https://gephi.org/

mais de um papel no mesmo filme, incluímos novos nós que deixam explícitos todos os papéis representados pelo ator. A Figura 11 representa o esquema do grafo nessa terceira etapa (**Modelagem 5**).

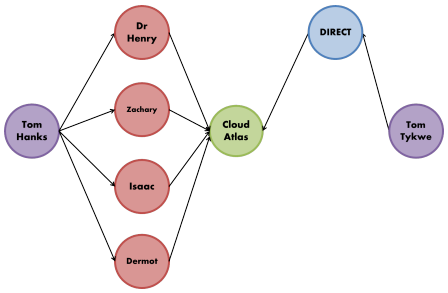

Figura 11. Esquema do grafo para a Modelagem 5

5.2.1. Análise dos Resultados. As Tabelas 2, 3 e 4 mostram os resultados obtidos para as medidas de centralidade proximidade, intermediação, autoridade e hub dos grafos construídos usando as três modelagens do experimento SMDB. Para o cálculo das centralidades, foram consideradas as opções de grafo não direcionado e direcionado. Os nós identificados nas tabelas correspondem aos 5 que tiveram maior valor de centralidade em cada caso.

Modelagem 3			
Não Direcionado		**Direcionado**	
Proximidade	**Intermediação**	**Autoridade**	**Hub**
Clint Eastwood	Tom Hanks	A few good men	Aaron Sorkin
Richard Harris	A few good men	Jerry Macguire	Al Pacino
Chris Columbus	Cloud Atlas	Speed Racer	Andy Wach.
Dina Meyer	Keanu Reeves	The Green Mile	Annabella Sci.
Ice-T	Jack Nicholson	Stand By Me	

Tabela 2. RESULTADOS DAS MEDIDAS DE CENTRALIDADE PARA MODELAGEM 3 NO EXPERIMENTO SMDB

Modelagem 4			
Não Direcionado		**Direcionado**	
Proximidade	**Intermediação**	**Autoridade**	**Hub**
Richard Harris	Tom Hanks	A Few Good Men	Aaron Sorkin
Clint Eastwood	A few good men	Jerry Macguire	Al Pacino
Chris Columbus	Cloud Atlas	Speed Racer	Andy Wach.
ACTED_IN	ACTED_IN	The Green Mile	
ACTED_IN	Keanu Reeves	Stand By Me	
DIRECTED	Jack Nicholson	A League Their Own	
Dina Meyer	The Green Mile	Cloud Atlas	

Tabela 3. RESULTADOS DAS MEDIDAS DE CENTRALIDADE PARA MODELAGEM 4 NO EXPERIMETO SMDB

O experimento apresentado nesta seção difere do experimento da seção 5.1 no que se refere à representação das entidades e relacionamentos no grafo, o que altera significativamente os resultados das medidas de centralidade.

Para as medidas de Proximidade e Intermediação, os três nós com maior valor de centralidade são os mesmos. Já nas modelagens 3 e 4 é possível notar uma mudança entre o segundo e terceiro colocados. Na modelagem 5, o nó que representa o filme *Cloud Atlas* toma a segunda posição mais alta no ranking de intermediação, que era ocupada na

Modelagem 5			
Não Direcionado		**Direcionado**	
Proximidade	**Intermediação**	**Autoridade**	**Hub**
Richard Harris	Tom Hanks	Cloud Atlas	Aaron Sorkin
Clint Eastwood	Cloud Atlas	A few good men	Al Pacino
Chris Columbus	A few good men	Jerry Macguire	Andy Wach
English Bob	Keanu Reeves	Speed Racer	
Bill Munny	The Green Mile	The green mile	
DIRECTED	Jack Nicholson	Stand By Me	
Dina Meyer	Hugo Weaving	A League Their Own	
Ice-T	"Paul Edgecomb"	Sleepless in Seattle	

Tabela 4. RESULTADOS DAS MEDIDAS DE CENTRALIDADE PARA MODELAGEM 5 NO EXPERIMENTO SMDB

modelagem 4 pelo nó *A Few Good Men*. Essa alteração se deve ao fato do filme *Cloud Atlas* possuir vários papéis representados pelos mesmos atores. Esse nó do tipo *Filme* também acaba sendo o de maior autoridade no grafo da modelagem 5; ou seja, vários outros nós apontam para esse *Filme*.

Outro nó que não estava presente no ranking de Intermediação para as modelagens 3 e 4 é o que representa o ator *Hugo Weaving*. Esse ator possui vários papéis em *Cloud Atlas* e está ligado a outros filmes no grafo. Portanto, o nó *Hugo Weaving* possui maior probabilidade de estar no caminho mínimo entre pares de nós, e por isso apresenta maior valor de intermediação na modelagem 5.

De forma geral, entende-se que os nós com maior valor de intermediação influenciam o valor dos nós ao redor. Assim, pode-se concluir que, o fato de considerar um relacionamento como nó no grafo não chega a ter um impacto tão grande, como verificado contrastando os resultados das modelagens 3 e 4.

No entanto, ao representar o atributo multivalorado do relacionamento entre duas entidades como um nó, isso faz com que as entidades envolvidas ganhem maior centralidade, em termos de intermediação e autoridade. Isso quer dizer, que se um nó já apresentasse um valor de centralidade significativo na modelagem 4, há um potencial aumento deste valor na modelagem 5, aumentando sua posição no ranking.

Em outras palavras, atributos de um relacionamento no esquema conceitual podem aumentar a importância das entidades nas análises, e por isso é importante avaliar se devem ser representados como nós o como arestas do grafo.

5.2.2. Heurística 2. A partir das observações mencionadas no experimento SMDB, foi obtida uma heurística que será apresentada a seguir. Dado um relacionamento R do tipo muitos-para-muitos $(n : m)$ entre 2 entidades E_1 e E_2, se um atributo P associado ao relacionamento R é relevante para a análise, então P deve ser modelado como um nó do grafo.

Ligado aos nós que representam as entidades E_1 e E_2 por meio de duas arestas R' e R''. Ou mais formalmente, como se segue.

Sejam:

- $G = (V, A)$; grafo
- $EC = (E, R, Pe, Pr)$; esquema conceitual

- $E = \{E_1, E_2, ...\}$; conjunto de entidades do esquema EC
- $R = \{R_1, R_2, ...\}$; conjunto de relacionamentos do esquema EC
- $E_i = \{e_i^1, e_i^2, ...\}$; conjunto de instâncias da entidade E_i
- $R_k = \{r_k^1, r_k^2, ...\}$; conjunto de instâncias do relacionamento R_k
- $r_k^z = (e_i^x, e_j^y)$; uma instância de R_k conecta um par de instâncias de entidades de E
- $Pe = \{Pe_1, Pe_2, ...\}$; conjunto de propriedades (atributos) de Entidades do esquema EC
- $Pr = \{Pr_1, Pr_2, ...\}$; conjunto de propriedades (atributos) de Relacionamentos do esquema EC
- $Pr_k = \{Pr_k^1, Pr_k^2, ...\}$; conjunto de propriedades (atributos) de um R_k do conjunto R do esquema EC
- $pr_k^{w,z} = (v_1, v_2, ...)$; conjunto de valores do atributo Pr_k^w de um relacionamento R_k, associado a uma instância de relacionamento r_k^z

Dado que:

- $E_1, E_2 \in E; e_1^x \in E_1; e_2^y \in E_2$;
- $R_1 \in R$;
- $r_1^z \in R_1; r_1^z = (e_1^x, e_2^y)$;
- $Pr_1^1 \in Pr_1$;
- $pr_1^{1,z} = (v_1, v_2, ..)$;

No mapeamento para o grafo G temos:

- $\forall e_1^x \in E_1, e_1^x \in V$;
- $\forall e_2^y \in E_2, e_2^y \in V$;
- Se Pr_1^1 é importante para o relacionamento R_1 e Pr_1^1 é multivalorado então
  $\forall r_1^z$ e $\forall v_s \in pr_1^{1,z}$: $v_s \in V$ e $(e_1^x, v_s), (v_s, e_2^y) \in A$

5.3. Experimento TMDB

Visando confirmar os resultados obtidos na seção 5.2, foram realizados novos experimentos utilizando um conjunto de dados maior, e o mesmo tipo de análises apresentadas anteriormente. O conjunto de dados utilizado nessa fase é uma base que contém dados sobre filmes e atores chamado TMDB[4]. Da base original do TMDB, foi extraído um subconjunto de dados usando o filtro gênero, "Crime" neste caso, e utilizando unicamente as informações sobre filmes e atores. Os experimentos dessa etapa reproduzem as mesmas três modelagens do experimento SMDB. Para garantir o cálculo das medidas de centralidade, foi considerado o maior componente conexo do grafo obtido para cada modelagem. O grafo gerado a partir da modelagem 3 possui 4486 nós e 5046 arestas. O grafo para a modelagem 4, 9532 nós e 10092 arestas. E o grafo da modelagem 5, 9561 nós e 10150 arestas. O objetivo desse experimento foi conferir se os comportamentos notados nos experimentos anteriores se repetiriam em um conjunto de dados maior. O fato do conjunto utilizado nesse experimento ser do mesmo domínio que no experimento SMDB, faz com que a expectativa de repetição do comportamento seja maior. Nesse experimento, as heurísticas até aqui obtidas não foram aplicadas, mas era esperado que o mesmo raciocínio para a geração dessas, fosse aqui usado.

O comportamento das medidas de centralidade para o novo conjunto de dados nas três modelagens foi semelhante ao observado no experimento SMDB. Em particular, observamos que as primeiras posições do ranking obtido em cada uma das três modelagens foi parecido ao obtido nos experimentos da seção 5.2.

As Tabelas 5, 6 e 7 mostram os 10 primeiros colocados no ranking obtido para cada uma das três modelagem usadas, considerando as medidas de centralidade proximidade, intermediação, hub e autoridade.

Assim, pode-se chegar nas mesmas conclusões:
(i) o fato de considerar um relacionamento como nó no grafo não chega a ter um impacto tão grande, como verificado contrastando os resultados das modelagens 3 e 4;
(ii) se um nó já apresentasse um valor de centralidade significativo na modelagem 4, há um potencial aumento deste valor na modelagem 5, aumentando sua posição no ranking.

6. Conclusão

Este trabalho abordou a tarefa de mapear dados relacionais para a representação em grafos, propondo um conjunto de heurísticas que visam auxiliar um usuário na tarefa de modelar o grafo. Em comparação com os trabalhos que abordam o mesmo problema, nosso trabalho difere por levar em consideração que um grafo pode seguir distintos modelos e modelagens. Além de levar em conta a análise topológica do grafo como fator para escolha da modelagem. Destacamos que, mesmo com o conjunto de coeficientes de análise em mãos, antes de aplicá-los, é necessário conhecer os esquemas conceitual e relacional do banco de dados relacional para melhor entender os dados e os resultados das medidas topológicas, para então alcançar uma boa modelagem do grafo.

Além das heurísticas abstraídas, os experimentos permitiram compreender como as métricas da teoria de grafos possuem interpretações diferentes. As interpretações estão relacionadas com a influência dos vértices no grafo. Em alguns momentos, as informações sobre as ligações e influências, extrapolavam a vizinhança de vértices, exigindo uma visão mais abrangente dos resultados das análises.

São reportados experimentos iniciais que gradativamente se tornavam mais complexos e permitiram identificar um conjunto de heurísticas. Esses experimentos foram realizados para análises topológicas, usando diferentes métricas em três conjuntos de dados: Zachary e SMDB, como conjuntos menores (*toy examples*), e TMDB, como conjunto de maior porte. No entanto, experimentos adicionais são necessários para validar e estender tais heurísticas. O uso de conjuntos de dados de diferentes domínios, de um conjunto maior de coeficientes analíticos e o uso de outros construtos de

4. http://www.themoviedb.org/

Modelagem 3			
Não Direcionado		**Direcionado**	
Proximidade	**Intermediação**	**Autoridade**	**Hub**
Dogville	Dogville	Angels & Demons	Angels & Demons
Ben Gazzara	Bloodline	Dogville	Dogville
Bloodline	Gert Fröbe	Armin Mueller-Stahl	Armin Mueller-Stahl
At Close Range	Ben Gazzara	Stellan Skarsgård	Stellan Skarsgård
The Big Sleep	Ten Little Indians	Dancer in the Dark	Dancer in the Dark
Lauren Bacall	The Big Sleep	Carmen Argenziano	Carmen Argenziano
Candy Clark	Humphrey Bogart	Tom Hanks	Tom Hanks
Patricia Clarkson	Vera Brühne	Elya Baskin	Elya Baskin
The Pledge	At Close Range	Ewan McGregor	Ewan McGregor
Nicole Kidman	Fritz Wepper	Thure Lindhardt	Thure Lindhardt

Tabela 5. RANKING DAS MEDIDAS DE CENTRALIDADE PARA A MODELAGEM 3 NO EXPERIMENTO TMDB

Modelagem 4			
Não Direcionado		**Direcionado**	
Proximidade	**Intermediação**	**Autoridade**	**Hub**
Dogville	Dogville	Straus	Angels & Demons
Jack McKay	Bloodline	Richter	Straus
Ma Ginger	Gert Fröbe	Silvano Ventivoglio	Richter
Ben Gazzara	Ben Gazzara	Cardinal	Silvano Ventivoglio
Vera	Rhys Williams	Robert Langdon	Cardinal
Grace Margaret Mulligan	Jack McKay	Carlo Ventresca	Robert Langdon
Rhys Williams	Ten Little Indians	Chartrand	Carlo Ventresca
Chuck	The Big Sleep	Mr. Gray	Chartrand
Bloodline	Humphrey Bogart	Father Simeon	Mr. Gray
Tom Edison Senior	Inspector Max Hornung	Swissguard	Father Simeon

Tabela 6. RANKING DAS MEDIDAS DE CENTRALIDADE PARA A MODELAGEM 4 NO EXPERIMENTO TMDB

Modelagem 5			
Não Direcionado		**Direcionado**	
Proximidade	**Intermediação**	**Autoridade**	**Hub**
Dogville	Dogville	Angels & Demons	Angels & Demons
Jack McKay	Bloodline	Straus	Straus
Ma Ginger	Gert Fröbe	Richter	Richter
Ben Gazzara	Ben Gazzara	Silvano Ventivoglio	Silvano Ventivoglio
Vera	Rhys Williams	Cardinal	Cardinal
Grace Margaret Mulligan	Jack McKay	Robert Langdon	Robert Langdon
Rhys Williams	Ten Little Indians	Carlo Ventresca	Carlo Ventresca
Chuck	The Big Sleep	Chartrand	Chartrand
Bloodline	Humphrey Bogart	Mr. Gray	Mr. Gray
Tom Edison Senior	Inspector Max Hornung	Father Simeon	Father Simeon

Tabela 7. RANKING DAS MEDIDAS DE CENTRALIDADE PARA A MODELAGEM 5 NO EXPERIMENTO TMDB

modelagem (e.g. relacionamentos n-ários), são planejados para experimentos futuros.

Agradecimentos

Este trabalho contou com o apoio parcial da CAPES (bolsa de estudos) e do CNPq (proc.307.647/2012-9).

Referências

[1] C. Bizer, "The d2rq mapping language," 2016, 14 nov. de 2016. [Online]. Available: http://d2rq.org/d2rq-language.html

[2] R. D. Virgilio, A. Maccioni, and R. Torlone, *Model-Driven Design of Graph Databases*. Cham: Springer International Publishing, 2014, pp. 172–185.

[3] D. Wardani and J. Küng, "Semantic mapping relational to graph model," in *Computer, Control, Informatics and Its Applications (IC3INA), 2014 International Conference on*, Oct 2014, intl, pp. 160–165.

[4] S. Bordoloi and B. Kalita, "Article: Designing graph database models from existing relational databases," *Int. J. of Computer Applications*, vol. 74, no. 1, pp. 25–31, July 2013.

[5] M. Newman, *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010.

[6] C. A. Heuser, *Projeto de banco de dados*. Sagra Luzzatto, 2001.

[7] P. P.-S. Chen, "The entity-relationship model—toward a unified view of data," *ACM Trans. Database Syst.*, vol. 1, no. 1, pp. 9–36, Mar. 1976.

[8] E. F. Codd, "A relational model of data for large shared data banks," *Commun. ACM*, pp. 377–387, 1970.

[9] M. A. Rodriguez and P. Neubauer, "Constructions from dots and lines," *Bulletin of the American Society for Information Science and Technology*, vol. 36, no. 6, pp. 35–41, 2010.

[10] R. D. Virgilio, A. Maccioni, and R. Torlone, "Converting relational to graph databases," in *First International Workshop on Graph Data Management Experiences and Systems*, ser. GRADES '13, no. 1. New York, NY, USA: ACM, 2013, intl, p. 6.

[11] K. Xirogiannopoulos, U. Khurana, and A. Deshpande, "Graphgen: Exploring interesting graphs in relational data," *PVLDB*, vol. 8, no. 12, pp. 2032–2035, 2015.

[12] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.