Uma Abordagem para Identificação de Entidades Influentes em Eventos Comentados nas Redes Sociais Online

Rayol M. Neto[1], **Bruno Á. Souza**[1], **Thais G. Almeida**[1], **Fabíola G. Nakamura**[1], **Eduardo F. Nakamura**[1]

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Manaus – AM – Brasil

{rayol,bruno.abia,tga,fabiola,nakamura}@icomp.ufam.edu.br

Abstract. *Online Social Networks (OSN) allow users to share content of any kind. In these networks, users can be considered as social sensors, where their opinions and comments about an event can be studied (e.g., characterizing events, recognizing influential entities). OSN can be modeled as complex networks, where the entities are represented by the vertices and the edges characterize the connection between them. Using this approach, we can determine which are the most influential entities in the studied context through centrality measures (e.g., Betweenness and Pagerank). Based on this, this article presents an analysis of the most influent entities from the perspective of an event in Online Social Networks and an study of the detected communities. As results we realize that the measures of centrality have different influential entities, this shows that although all the entities are part of the same event, some have greater influence depending on the studied aspect. Regarding the community detection we notice that entities that were grouped together, are directly linked to a sub-event within the studied context.*

Resumo. *Redes Sociais Online (RSO) permitem aos usuários compartilhar conteúdo de qualquer tipo. Nestas redes, os usuários podem ser vistos como sensores sociais, onde suas opiniões e comentários a respeito de um evento podem ser utilizados para estudo (e.g., caracterização de eventos, reconhecimento de entidades influentes). RSO podem ser modeladas como redes complexas, onde as entidades são representadas pelos vértices e as arestas caracterizam a conexão entra elas. Utilizando esta abordagem, podemos determinar quais são as entidades mais influentes no contexto estudado através de medidas de centralidade (e.g., Betweenness e Pagerank). Com base nisto, este artigo apresenta uma análise das entidades mais fluentes sob a perspectiva de um evento em Redes Sociais Online e apresenta um estudo das comunidades detectadas. Como resultados percebemos que as medidas de centralidade apresentam diferentes entidades influentes, isto demonstra que embora todas as entidades façam parte do mesmo evento, algumas tem influência maior dependendo do aspecto estudado. Com relação a detecção de comunidades percebemos que entidades que foram agrupadas juntas, estão diretamente ligadas a um sub-evento dentro do contexto estudado.*

1. Introdução

Com a disseminação da Internet, as Redes Sociais Online (RSO) emergiram possibilitando a conexão de pessoas através de ambientes virtuais, que disponibilizam recursos

como compartilhamento de fotos, vídeos, opiniões e percepções pessoais [Atefeh and Khreich 2015; Benevenuto et al. 2011]. Atualmente, essas RSO fazem parte do cotidiano de pessoas e empresas (e.g., Twitter possui 313 milhões de usuários mensais[1]), fornecendo recursos cada vez mais atrativos aos seus usuários. Com o aumento da conectividade, as RSO podem ser utilizadas como fontes de dados para pesquisas, pois possuem uma elevada quantidade de dados produzidas pelos seus usuários [Souza et al. 2016].

Partindo desse princípio, cada usuário nesses ambientes pode ser visto como um sensor colaborativo, onde suas opiniões e percepções a respeito de um determinado acontecimento podem ser utilizadas para caracterizar eventos [Sakaki et al. 2010], reconhecer entidades influentes [Cha et al. 2010] e analisar sentimentos [Almeida et al. 2016]. Nesse contexto, técnicas de aprendizagem de máquina supervisionadas e não-supervisionadas são possíveis soluções para extrair estas informações [Souza et al. 2016; Petrović et al. 2010]. Outra abordagem que também pode ser utilizada são as Redes Complexas, onde as interações de usuários ou informações compartilhadas podem ser modeladas através de arestas, e as entidades representadas por nós, que em conjunto formam a rede [Mislove et al. 2007]. Essa estratégia tem sido amplamente utilizada para reconhecimento de perfis influentes e principais entidades, disponibilizando métodos e métricas para tais finalidades.

A partir desse cenário, existem trabalhos recentes de Redes Complexas que foram capazes de definir as entidades mais influentes de acordo com diferentes métricas de centralidade [Rostami and Mondani 2015; Beveridge and Shan 2016]. Outros autores utilizaram essas métricas no contexto de Redes Sociais Online [Mislove et al. 2007; Oliveira et al. 2015] e estudaram sua estrutura e evolução [Kumar et al. 2010]. Entretanto, as soluções existentes não apresentam uma abordagem que combine a extração de entidades em textos, verificação das relações entre entidades e a representação através de Redes Complexas para reconhecimento de principais entidades dentro de um evento comentado nas mídias sociais.

Com isso, as principais contribuições deste trabalho são: (i) uma abordagem para modelar redes complexas a partir de comentários compartilhados nas RSO; (ii) a análise que demonstra o impacto de diferentes métricas de centralidade aplicadas nesse contexto, a fim de identificar as entidades mais influentes (entidades que possuem maior relevância no contexto estudado) em um evento. Para validação da abordagem proposta, realizamos um estudo de caso considerando as 15 fases da Operação Lava Jato em 2016.

Este artigo está organizado da seguinte maneira. Na Seção 2 são discutidos os trabalhos relacionados. Em seguida, a Seção 3 apresenta a abordagem proposta. Os resultados são demonstrados na Seção 4. Por fim, a Seção 5 apresenta as considerações finais e direções futuras.

2. Trabalhos Relacionados

Medidas de centralidade foram criadas como forma de medir a importância de um nó em uma rede [Newman 2003]. Por possuírem essa característica, essas métricas vêm sendo empregadas em redes sociais [Jamali and Abolhassani 2006] com o intuito de identificar quais entidades são as mais influentes.

[1]`https://about.twitter.com/company`

No trabalho de Rostami and Mondani [2015], os autores apresentam um estudo para saber se diferentes bases de dados a respeito do mesmo evento influenciam na análise dos resultados das Redes Complexas. Eles utilizam três bases de dados diferentes sobre uma gangue de rua Sueca para criar as Redes Complexas, e comparam essas redes calculando as métricas de distância, centralidade e agrupamento. No cálculo das distâncias foi utilizado o *Graph Edit Distance* (*GED*), para calcular a centralidade utilizam *Degree centrality* (*Dc*) e *Betweenness centrality* (*Bc*) e por fim para calcular o agrupamento utilizam o *local Clustering Coefficient* (C_i). Como resultado, os autores demonstram claramente que diferentes bases de dados para o mesmo fenômeno produzem cenários diferentes.

Oliveira et al. [2015] apresentam um estudo sobre o parlamento brasileiro no Twitter. Para a análise dos dados, utilizam as métricas: grau, assortatividade, HITS, *Betweenness*, coeficiente de clusterização, reciprocidade, diâmetro e grau de distância. Os autores levam em consideração que os vínculos sociais ou políticos foram expressos pelas suas conexões e frequência de interação entre os usuários no Twitter. Os resultados da análise das medidas de centralidade apontam que existem mais parlamentares sendo seguidos do que seguindo outros usuários, a rede é densa e de pequeno diâmetro, além de possuir um alto coeficiente de clusterização. Adicionalmente, foi observado que há uma grande reciprocidade nas ligações e os parlamentares com os maiores valores de *Betweenness* exercem algum tipo de liderança no parlamento Brasileiro.

Beveridge and Shan [2016] apresentam o impacto da análise de diferentes métricas (*Degree*, *Weighted degree*, *Eigenvector*, *PageRank*, *Closeness* e *Betweenness*) de centralidade em um mesmo domínio. No artigo, foi criada uma rede complexa onde os nós são os personagens que aparecem no terceiro livro do universo de *Game of Thrones* e as arestas a relação entre esses personagens. Os resultados apresentados mostram que dependendo da medida de centralidade utilizada, um personagem pode ser mais ou menos influente no enredo estudado.

Mislove et al. [2007] apresentam um estudo de medição em larga escala e análise da estrutura de múltiplas RSO, onde foram coletados dados de 4 mídias sociais Flickr, YouTube, LiveJournal e Orkut. Foi observado que os nós dos usuários na medida *Indegree* tendem a igualar os da medida *Outdegree* e que as redes contêm um núcleo densamente conectado de nós de alto grau. Além disso, percebeu-se que este núcleo liga pequenos grupos de nós de baixo grau fortemente agrupados nas margens da rede.

Na pesquisa de Liu et al. [2016], os autores relacionam o sucesso de um clube de futebol com suas atividades no mercado de transferência de jogadores de uma perspectiva de Redes Complexas. Nessa rede, os nós são os clubes de elite, e as arestas direcionadas que conectam os nós são as transferências de jogadores e os empréstimos. Particularmente, estuda-se a relação entre as propriedades do nó e as funcionalidades dos clubes profissionais. Os autores utilizam 5 medidas de centralidade *Eigenvector*, *PageRank*, *Effective size*, *Betweenness* e *Closeness*. Os resultados mostram que o desempenho e a rentabilidade dos clubes no mercado de transferência estão fortemente associados às propriedades de centralidade de seus nós correspondentes na rede de transferência de jogadores.

Os principais diferenciais da abordagem proposta neste trabalho, quando comparado às soluções apresentadas, são: (i) extrair entidades do evento de forma automatizada;

(ii) avaliar diferentes medidas de centralidade a partir de dados de Redes Sociais Online; (iii) identificar as entidades mais influentes em um evento comentado em mídias sociais.

3. Abordagem Proposta

A abordagem proposta neste trabalho consiste das seguintes etapas (ilustradas na Figura 1): (i) coleta de dados provenientes das Redes Sociais Online para a construção da base dados; (ii) pré-processamento dos dados, a fim de retirar ruídos existentes na base de dados; (iii) reconhecimento das entidades nomeadas; (iv) cálculo das ocorrências conjuntas de entidades presentes no mesmo documento, para a geração da rede complexa; e (v) análise das comunidades e principais entidades da rede complexa modelada.

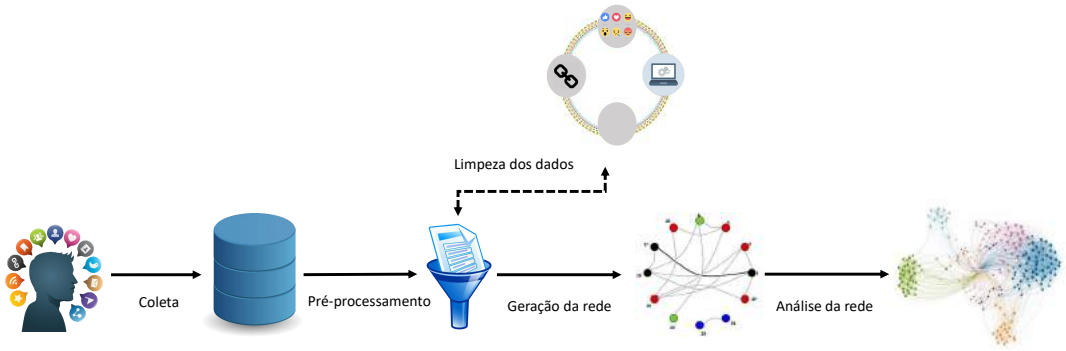

Figura 1. Etapas da abordagem proposta.

3.1. Coleta e Pré-processamento

A abordagem proposta se inicia na fase de coleta, onde os dados podem ser extraídos de uma RSO, a fim de gerar a base de dados. Após a fase de coleta de dados e construção da base de dados, o pré-processamento é necessário para eliminar os ruídos da base de dados. Nessa etapa removemos *urls* e menções.

3.2. Criação da Rede (Grafo)

Nessa fase é utilizada a abordagem de reconhecimento de entidade nomeada (NER - *Named Entity Recognition*) para extrair quais entidades estão sendo mencionadas na base (conforme ilustrado na Figura 2). Desta forma, o dados passam a ser representados somente pelas entidades que foram detectadas, pois assim podemos fornecer as entradas necessárias para a geração da rede. As entidades representam os vértices do grafo e o peso das arestas equivale à quantidade de vezes que a entidade a é mencionada junto com a entidade b ao longo da base de dados. Calculamos as ocorrências conjuntas através da Equação 1, onde assumimos que a frequência dos termos $F(a, b)$ é soma das frequências de ocorrências dos termos em todos os documentos.

$$F_{(a,b)} = \sum_{j=1}^{D} f_{((a,b),j)} \tag{1}$$

Nessa etapa, consideramos que relações entre as entidades são bidirecionais, ou seja, tanto as frequências $F(a, b)$ como $F(b, a)$, são representadas como uma única aresta

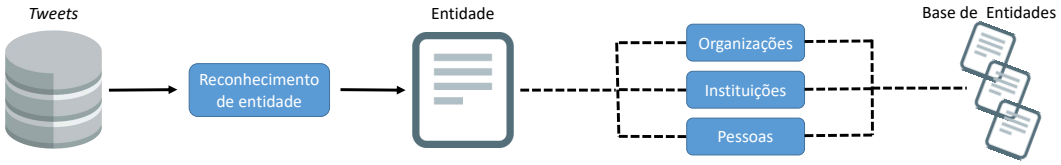

Figura 2. Estrutura utilizada no reconhecimento de entidade.

na rede (grafo) e com isso, poderíamos somar os valores das frequências, como por exemplo, as entidades "X" e "Y" tem frequência igual a 20, e a inversão "Y" e "X" tem frequência igual a 30, com isso, ao final teríamos uma única relação entre as duas entidades com peso igual a 50.

3.3. Análise da Rede

Na fase de análise da rede, aplicamos 6 métricas de centralidade (*Degree*, *Weighted Degree*, *Eigenvector*, *Pagerank*, *Closeness* e *Betweenness*), com o intuito de estudar o impacto dessas métricas sobre uma rede baseada em opiniões pessoais de usuários de RSO. Adicionalmente, aplicamos o algoritmo de *Louvain* [Blondel et al. 2008] para detectar comunidades.

As métricas de centralidade podem ser utilizadas para identificar vértices importantes e seu nível de influência dentro de uma rede [Beveridge and Shan 2016]. Cada métrica leva em consideração diferentes características da rede para mensurar a importância dos vértices. O *Degree Centrality* considera o número de arestas incidentes que um vértice possui [Freeman 1978]. O *Weighted Degree Centrality* leva em consideração o valor do peso das arestas, ou seja, seu cálculo é feito a partir da soma dos pesos das arestas incidentes em um vértice [Barrat et al. 2004]. Na equação abaixo, V corresponde ao conjunto de vértices do grafo:

$$v_i = \sum_{k \in V} a_{v_i, v_k}. \tag{2}$$

O *Eigenvector Centrality* assume que um vértice importante está conectado com outros nós importantes, onde a importância x_v de um vértice i pode ser medido pelo somatório da importância dos seus vizinhos (equação 3). Dado uma matriz de vértices adjacentes A com entradas $a_{i,k}$ e $\lambda \neq 0$ sendo um autovalor da matriz A, este cálculo é aplicado sobre cada vértice da rede, não sendo necessário ter várias arestas conectadas a um vértice para se obter um alto *eigenvector* [Newman 2008].

$$v_i = \frac{1}{\lambda} \sum_{k} a_{k,i} v_k \tag{3}$$

Diferente do *Eigenvector*, o *PageRank Centrality* assume que a contribuição de centralidade dos vértices não é a mesma e deve ser penalizada proporcionalmente de acordo com a quantidade de vizinhos, ou seja, se um vértice com alto grau possuir muitos vizinhos para dividir sua importância, sua centralidade será mais penalizada que os outros nós com menos vizinhos. O *PageRank Centrality* v_i de um vértice i é calculado conforme

equação 4, onde $a_{k,i}$ é a matriz de vértices adjacentes do nó a qual aplicamos a equação, d_k é o grau de saída do vértice k, α e β são constantes e v_k é o valor de *pagerank* do vértice vizinho [Brandes and Erlebach 2005].

$$v_i = \alpha \sum_k \frac{a_{k,i}}{d_k} v_k + \beta \tag{4}$$

O *Closeness Centrality* calcula a distância de um nó para todos os vértices da rede como forma de medir a importância deste vértice. Essa métrica é aplicada conforme a equação 5, onde $d(y, x)$ é a distância entre x e y [Rochat 2009].

$$C(v) = \frac{1}{\sum_y d(y, x)} \tag{5}$$

O *Betweenness Centrality* é uma medida de centralidade que no decorrer do seu cálculo leva em consideração os caminhos mínimos. Matematicamente, podemos obter o *betweenness* v_i de um vértice i através da equação 6, onde σ_{st} é o valor total de caminhos mínimos entre os vértices s e t, e $\sigma_{st}(i)$ é quantidade de caminhos mínimos que passam pelo vértice i [Brandes 2001].

$$v_i = \sum_{s \neq t \neq z} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{6}$$

Com relação à detecção de comunidade, utilizamos a modularidade, uma função que mede a qualidade da divisão de uma rede em grupos ou comunidades [Newman 2006]. Neste trabalho aplicamos o método heurístico proposto por Blondel et al. [2008], chamado de algoritmo de *Louvain*, que se baseia na maximização da modularidade. Este algoritmo guloso é composto de duas partes, na primeira é designada uma comunidade diferente para cada nó i da rede, ou seja, o número de comunidades é equivalente ao número de nós. Na segunda parte, para cada nó i, leva-se em consideração todos os seus vizinhos. Nessa etapa é avaliado o ganho de modularidade de i, ao mudá-lo para a comunidade de seus vizinhos j. O nó i fica na comunidade onde obtiver o maior ganho de modularidade. Matematicamente, este método pode ser definido por:

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m} \right)^2 \right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right], \tag{7}$$

onde $\sum_{in}$ é a soma dos pesos das arestas na comunidade C, $\sum_{tot}$ é a soma dos pesos das arestas incidentes nos nós em C, k_i é a soma dos pesos das arestas incidentes no nó i, $k_{i,in}$ é a soma dos pesos das arestas partindo do nó i para todos os nós em C e m é a soma dos pesos de todas as arestas da rede. Todo esse processo é aplicado repetidas vezes, quando a modularidade máxima é alcançada, ou seja, nenhum nó i pode aumentar sua modularidade, o processo é interrompido.

4. Avaliação da Abordagem Proposta

Essa Seção apresenta uma análise da influência das entidades quando aplicadas diferentes métricas de centralidade e um estudo sobre as comunidades geradas a partir do algoritmo

de *Louvain*. Adicionalmente, detalhamos a metodologia utilizada na aplicação da abordagem proposta.

4.1. Metodologia

Neste trabalho, criamos uma base de dados a partir de informações coletados do Twitter, onde utilizamos a *Search API* do Twitter, que permite a coleta de dados históricos publicados pelos usuários. Para filtrar dados referentes ao evento em estudo, utilizamos um recurso disponível pela API para filtrar informações de acordo com palavras-chave.

Nesse contexto, a nossa coleta de *tweets* foi feita utilizando a palavra-chave "Lava Jato", uma vez que nossa base de dados é relativa a Operação Lava-Jato, deflagrada em 2014 pela Polícia Federal do Brasil, cujo objetivo consiste em investigar a prática de crimes financeiros e desvio de recursos públicos. Vale ressaltar, que a coleta é relativa ao período de 01/01/2016 à 20/11/2016, pois consideramos somente as fases realizadas em 2016, ou seja, da 22ª à 36ª fase.

Ao longo da criação da base, observamos que existiam *tweets* que apresentavam o termo "Lava Jato", porém não estavam conectados aos acontecimentos do evento estudado, como por exemplo, *chekins* feitos em postos de lavagem. Com isso, eliminamos esses documentos que não estavam relatando opiniões ou notícias sobre as fases da Operação Lava Jato. Ao final, nossa base consiste em 652.210 *tweets* escritos em português que mencionavam o termo "Lava Jato" e tinham ligação ao contexto em estudo.

Após a geração da base de dados, realizamos o pré-processamento, onde removemos as *urls* e menções dos *tweets* e utilizamos o PolyGlot-NER[2] [Al-Rfou et al. 2015] para reconhecimento de entidades nomeadas, pois apresenta suporte para o idioma Português em tarefas de processamento de linguagem natural.

Para a construção do grafo, utilizamos a ferramenta Gephi[3], que disponibiliza recursos de *plot* de redes, assim como detecção de comunidade. Para o cálculo das medidas de centralidade, utilizamos a ferramenta graph-tool[4]. Por fim, é valido ressaltar que para a geração da rede estudada, consideramos apenas *tweets* que possuíam duas ou mais entidades citadas ao longo do texto, além disso, as entidades que representavam componentes desconexos do grafo foram descartadas, uma vez que não possuíam informações pertinentes. Totalizando assim, 59 entidades detectadas que atendiam essa condição.

4.2. Avaliação das Medidas de Centralidade

As medidas de centralidade selecionadas representam diferentes aspectos da rede, como nós vizinhos, visão geral da rede e arestas incidentes. Com isso, comparamos entre si as medidas que possuem os mesmos aspectos (e.g., *Eigenvector* e *Pagerank*). Com o intuito de apresentar as mesmas entidades nos resultados, selecionamos as 15 entidades mais influentes de cada medida e depois fizemos a interseção entra elas, ou seja, apresentamos aqui somente as entidades mais influentes que apereceram em todas as 6 métricas utilizadas, são elas: "Dilma", "Eduardo Cunha", "Lula", "Michel Temer", "Odebrecht", "PF", "PT", "Sérgio Moro"e "STF".

[2] `http://polyglot.readthedocs.io`
[3] `https://gephi.org/`
[4] `https://graph-tool.skewed.de`

A Figura 3 apresenta as medidas de centralidade de todas as métricas analisadas, vale ressaltar que as medidas foram normalizadas com valores entre 0 e 1. Para fins de análise, as entidades que aparecem desconexas da rede principal, foram tratadas como ruídos e consequentemente excluídas.

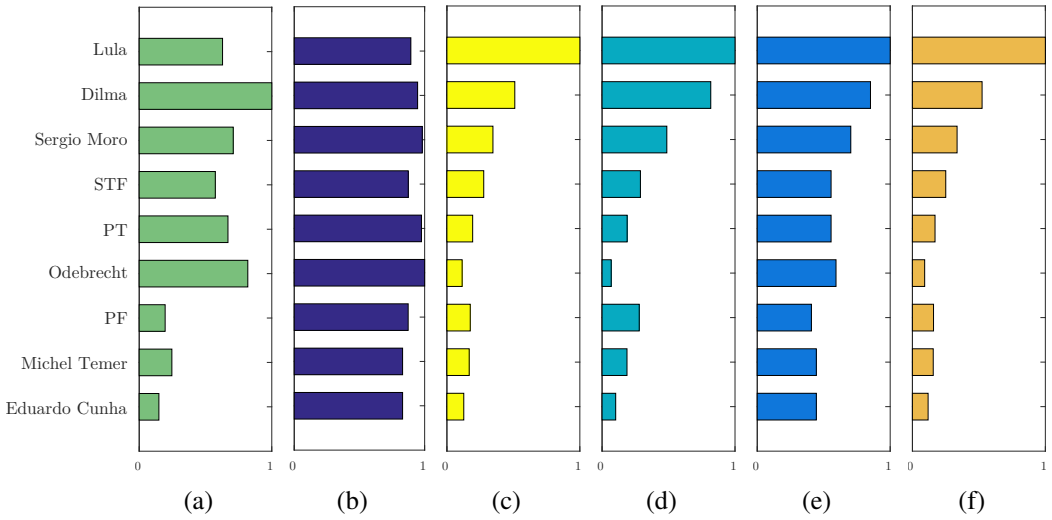

Figura 3. As 9 entidades mais influentes dentre as medidas de centralidade, normalizadas entre 0 e 1. (a) Betweenness, (b) Closeness (c) Pagerank, (d) Eigenvector, (e) Degree, (f) Weighted Degree.

Betweenness e *Closeness*

Primeiro comparamos duas métricas de centralidade que dão uma visão geral da rede, são elas *Closeness* e *Betweenness*. A *Closeness* apresenta a distância média de um vértice para todos os outros da rede. Nessa medida, os valores de todas as entidades ficaram bem próximos, conforme apresentado na Figura 3(b). Isso demonstra que todas as entidades estão relativamente próximas uma das outras, pois todas elas fazem parte do mesmo evento. A entidade que apresenta maior *Closeness* é a "Odebrecht", pois a sua distância média é a menor em relação a todas as outras entidades, uma vez que a empreiteira é uma das entidades mais citadas na operação Lava Jato.

Conforme apresentado na Seção 3 o *Betweenness* mede a frequência que um vértice i aparece nos caminhos mínimos entre os vértices s e t. Assim, percebemos por essa medida que a entidade "Dilma" é a mais influente, logo podemos inferir que é a entidade que possui uma ligação mais central em relação aos outros, ou seja, para os usuários do Twitter, a entidade "Dilma" é a que possui maior relacionamento com as outras entidades da Lava Jato. Outro fator a se observar, é que a entidade "Odebrecht" aparece como segundo nó mais influente de acordo com essa medida de centralidade, uma vez que muitos dos escândalos relacionados à operação Lava Jato estão relacionados à empreiteira, logo as entidades envolvidas têm conexão com a empresa.

Pagerank e *Eigenvector*

Neste trabalho, comparamos *Pagerank* e *Eigenvector*, pois conforme apresentado na Seção 3, eles diferem apenas na contribuição da centralidade dos vértices. Como pode ser observado nas Figuras 3(c) e 3(d), tanto o *Eigenvector* quanto o *Pagerank* possuem resultados próximos, ou seja, a contribuição da centralidade dos vértices faz com que as entidades tenham importâncias diferentes na rede. Porém a ordem de influência é pouco alterada, pois observamos que a diferença de ambos aparece na quinta entidade mais influente.

No *Eigenvector* a entidade "PF" possui mais influência que a entidade "PT", enquanto no *Pagerank* é o contrário. Como o *Pagerank* assume que a contribuição de centralidade dos vértices não é a mesma, a medida penaliza a entidade de acordo com a quantidade de nós vizinhos. Percebemos então que o número de nós ligados a essas entidades influenciou na importância do nó em questão. Analisando os dados, enquanto a entidade "PT" possui 14 nós adjacentes, a entidade "PF" possui apenas 7.

Com relação as entidades mais influentes, "Lula" e "Dilma" se destacam em ambas as medidas de centralidade. Na rede estudada, os nós adjacentes dessas entidades também possuem uma medida alta de influência, além disso, são os nós que possuem mais vizinhos respectivamente. Com isso, percebemos que ambos são nós centrais da rede, aparecendo com o maior número de conexões. Logo, conforme os *tweets* dos usuários, as entidades "Lula" e "Dilma" são as mais influentes na operação deflagrada pela Polícia Federal, pois estão altamente conectados com outras entidades influentes.

Degree e *Weighted Degree*

Degree e *Weighted Degree* são analisadas juntos pois são duas medidas de centralidade que levam em consideração somente as arestas incidentes no vértice em questão. Do mesmo modo que *Pagerank* e *Eigenvector*, as Figuras 3(e) e 3(f) ilustram "Lula" e "Dilma" como as entidades mais influentes, pois ambas possuem o maior número de arestas incidentes, além do peso de suas arestas serem maiores que as das demais entidades, ou seja, são as entidades que mais possuem ligações com outras entidades na Lava Jato, além disso essas ligações são as que mais se repetem na rede.

Porém, é válido ressaltar a diferença de influência da entidade "Odebrecht". Enquanto na medida *Degree* é a quarta entidade mais importante, no *Weighted Degree* é a nona mais influente, como ilustra a Figura 3(f). Isto é explicado pelo fato de que por mais que a "Odebrecht" possua várias arestas incidentes, o peso das mesmas é baixo, o que a torna pouco influente nessa métrica.

4.3. Avaliação de Comunidades

A Figura 4 ilustra a rede social modelada a partir do contexto estudado. No grafo apresentado, cada comunidade é representada por uma cor distinta. O tamanho dos vértices corresponde ao seu valor de *PageRank*, enquanto tamanho de seu rótulo (nome) representa a medida de centralidade *Betweenness*. A espessura das arestas representa seu peso.

Na rede estudada foram detectadas 5 comunidades (Figura 4). Porém, por se tratar de um algoritmo guloso, dependendo do nó de origem escolhido, o número de

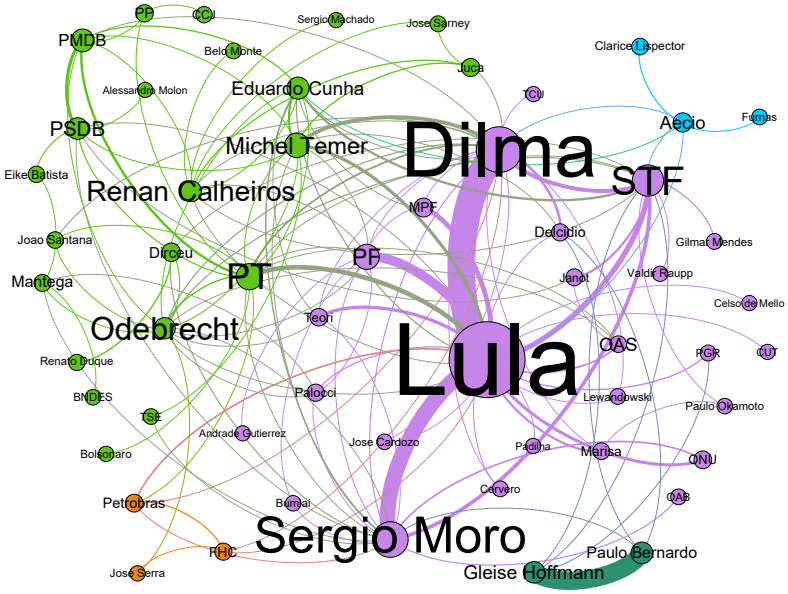

Figura 4. Rede social gerada a partir dos *tweets* dos usuários em relação a operação Lava Jato. A cor do vértice indica sua comunidade. O tamanho do vértice corresponde ao seu valor *PageRank*, e o tamanho do seu rótulo corresponde ao seu *Betweenness*. A espessura das arestas representa seu peso.

comunidades pode mudar. É importante ressaltar que o algoritmo de *Louvain* escolhe o nó de origem de forma aleatória.

A comunidade verde escuro é composta pelas entidades "Gleisi Hoffmann" e "Paulo Bernardo", ambas as entidades foram agrupadas nessa comunidade por terem sido indiciadas juntos como réus da Operação Lava Jato, além disso, são marido e mulher. Nos *tweets* relacionados a este assunto, essas entidades sempre apareciam juntas, como mostra o peso da aresta na Figura 4, representado por sua espessura.

Algumas comunidades apresentaram ruídos, como no caso da azul claro, percebemos que a entidade "Clarisse Lispector" não é uma das personagens da operação Lava Jato, porém em alguns casos quando os usuários do Twitter se referiam a entidade "Aécio", eles o relacionavam com a famosa escritora "Clarice Lispector", um exemplo de *tweet* é: *"Aécio já é mais citado na Lava Jato que Clarice Lispector no Facebook"*.

As maiores comunidades apresentadas (verde e roxo), têm este tamanho pois utilizando o algoritmo de *Louvain*, a modularidade é maximizada de forma que os agrupamentos tenham uma densidade interna entre os nós maior do que a densidade entre os nós que interligam as comunidades vizinhas, ou seja, a modularidade maximizada apresenta os nós que possuem modularidade alta entre si, criando assim uma comunidade. Isto implica que as entidades agrupadas nessas comunidades estão ligadas dentro de um sub-evento no contexto estudado (neste artigo, caracterizamos sub-evento com um evento originado a partir de um outro evento). Por exemplo, as entidades "Lula" e "Sergio Moro" estão na mesma comunidade, pois ambas estão relacionadas a um mesmo sub-evento. No âmbito da operação Lava Jato, a entidade "Sergio Moro" concedeu um mandado de condução coercitiva para que a entidade "Lula" prestasse depoimento nas investigações.

Com relação a comunidade laranja, percebemos que as entidades também estão relacionadas a um sub-evento, como pode ser observado nos *tweets* “Lava-Jato parada na Corte cooptado: Teori envia para Moro suspeita de propina na Petrobras no governo FHC” e “José Serra é citado em negociação de delação premiada da OAS na Lava Jato #epoca #exame #recordtv #moro #fhc #psdb”. Ambos extraídos da base de dados coletada.

4.4. Discussão

Percebemos que os dados coletados a partir de sensores colaborativos (usuários de redes sociais), servem para detecção de entidades influentes a partir de medidas de centralidade. Embora não haja uma medida de centralidade geral, Como pode ser observado na Tabela 1, cada medida fornece informações complementares sobre a rede, dependendo do que está sendo analisado, ou seja, analisá-las em conjunto apresenta diferentes aspectos da rede e de seus personagens.

Tabela 1. Medidas de Centralidade e suas entidades mais influentes.

Medidas de Centralidade	Aspecto da Métrica	Entidades mais influentes
Betweenness	Caminhos entre nós	Dilma, Odebrecht, Sergio Moro PT e Lula
Closeness	Caminhos entre nós	Odebrecht, Sergio Moro, PT Dilma e Lula
Eigenvector	Nós vizinhos	Lula, Dilma, Sergio Moro STF e PF
Pagerank	Nós vizinhos	Lula, Dilma, Sergio Moro STF e PT
Degree	Arestas incidentes	Lula, Dilma, Sergio Moro Odebrecht e STF
Weighted Degree	Arestas incidentes	Lula, Dilma, Sergio Moro STF e PT

Podemos perceber, conforme apresentado na Tabela 1, que *Closeness* e *Betweenness* foram as duas únicas medidas de centralidade que divergiram das outras, no que se refere a entidade mais influente, pois dentre as medidas de centralidade utilizadas neste artigo, são as únicas que estabelecem uma relação do caminho dos nós percorrido entre dois vértices. Com relação aos dados coletados sobre a operação Lava Jato, demonstramos que a entidade “Odebrecht” é o nó central da rede, ou seja, está diretamente conectada a várias entidades, uma vez que é uma das empresas investigadas na operação. Além disso, demonstramos também que a entidade “Dilma” é o maior elo de ligação entre as entidades, pois na época era a presidente do Brasil e seu nome estava ligado à várias entidades (pessoas ou organizações).

As entidades “Lula”, “Dilma” e “Sérgio Moro”, por outro lado, aparecem respectivamente como as entidades mais influentes no *Eingenvector*, *Pagerank*, *Degree* e *Weighted Degree*. Isto acontece devido ao alto grau de conexão destas entidades com os seus nós vizinhos, ou seja, além das entidades possuírem as maiores quantidades de vizinhos, suas arestas possuem pesos maiores que os outros. Como estas medidas levam em consideração essas características, essas entidades apresentaram ser as mais influentes da rede. No evento estudado, demonstramos que estas 3 entidades são as mais influentes,

possuindo uma forte ligação entre elas e entre os nós vizinhos, demonstrando assim que são peças chaves da operação Lava Jato.

Em relação à detecção de comunidades, observamos que o algoritmo de *Louvain* é uma abordagem viável a ser aplicada ao contexto de Redes Sociais Online. Ainda mais, demonstramos que as comunidades geradas agruparam as entidades que possuem ligação dentro de um mesmo sub-evento. Onde cada comunidade apresenta um cenário diferente que pode ser estudado em particular.

5. Conclusão

Neste trabalho apresentamos uma análise de entidades influentes sob a perspectiva dos comentários de usuários a respeito de um determinado evento em Redes Sociais Online, para isto utilizamos medidas de centralidade com o intuito de identificar quais são essas entidades influentes, além disso separamos as entidades de acordo com sua comunidade, utilizando o algoritmo de *Louvain*.

Os resultados apresentados demonstram que o pré-processamento automatizado proposto, é viável para modelar a rede. Com relação a medidas de centralidade, percebemos que não há uma métrica geral a ser aplicada no contexto de Redes Sociais Online, como cada medida fornece diferentes informações sobre a rede, elas se complementam. Percebemos que dependendo da medida de centralidade estudada, as entidades mais influentes variaram de acordo com o aspecto da métrica, isto demonstra que embora todas as entidades façam parte do mesmo evento, algumas tem influência maior dependendo do que está sendo estudado. No que diz respeito a detecção de comunidades, percebemos que o algoritmo de *Louvain* agrupou as entidades que estão diretamente ligadas a um sub-evento dentro do contexto da Operação Lava Jato.

Como trabalhos futuros podem ser utilizadas outras bases sobre o mesmo evento para comparação se há diferença de influência entre as entidades de acordo com a base estudada, além disso pode-se aplicar outras medidas de centralidade para analisar outros aspectos da rede e das relações entre as entidades.

Referências

Al-Rfou, R., Kulkarni, V., Perozzi, B., and Skiena, S. (2015). Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594. SIAM.

Almeida, T. G., Souza, B. A., Menezes, A. A., Figueiredo, C. M., and Nakamura, E. F. (2016). Sentiment analysis of Portuguese comments from Foursquare. In *Proceedings of the 22Nd Brazilian Symposium on Multimedia and the Web*, pages 355–358. ACM.

Atefeh, F. and Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164.

Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752.

Benevenuto, F. R., Silveira, D., Bombonato, L., Fortes, R., and Pereira Junior, Á. R. (2011). Entendendo a twitteresfera brasileira. In *Proceedings of the 8th Simpósio Brasileiro de Sistemas Colaborativos*.

Beveridge, A. and Shan, J. (2016). Network of thrones. *Math Horizons*, pages 18–22.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2):163–177.

Brandes, U. and Erlebach, T. (2005). *Network analysis: methodological foundations*, volume 3418. Springer Science & Business Media.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the 4th International AAAI Conference On Web and Social Media*, pages 335–338. AAAI.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239.

Jamali, M. and Abolhassani, H. (2006). Different aspects of social network analysis. In *Proceedings of the International Conference on Web Intelligence*, pages 66–72. IEEE.

Kumar, R., Novak, J., and Tomkins, A. (2010). Structure and evolution of online social networks. In *Link mining: models, algorithms, and applications*, pages 337–357. Springer.

Liu, X. F., Liu, Y.-L., Lu, X.-H., Wang, Q.-X., and Wang, T.-X. (2016). The anatomy of the global football player transfer network: Club functionalities versus network properties. *PloS One*, 11(6):e0156504.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings of the 7th of Internet Measurement Conference*, pages 29–42. ACM.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582.

Newman, M. E. (2008). The mathematics of networks. *The new palgrave encyclopedia of economics*, 2(2008):1–12.

Oliveira, L. S., Amaral, M. S., and Aguiar, I. S. (2015). O parlamento brasileiro no twitter: uma análise de rede social. In *Proceedings of the 12th Simpósio Brasileiro de Sistemas Colaborativos*, pages 138–143.

Petrović, S., Osborne, M., and Lavrenko, V. (2010). Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189. Association for Computational Linguistics.

Rochat, Y. (2009). Closeness centrality extended to unconnected graphs: The harmonic centrality index. In *Proceedings of the 6th International Conference on Applications of Social Network Analysis*.

Rostami, A. and Mondani, H. (2015). The complexity of crime network data: A case study of its consequences for crime control and the study of networks. *PloS one*, 10(3):e0119309.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860. ACM.

Souza, B. A., Almeida, T. G., Menezes, A. A., Nakamura, F. G., Figueiredo, C. M., and Nakamura, E. F. (2016). For or against?: Polarity analysis in tweets about impeach-

ment process of brazil president. In *Proceedings of the 22Nd Brazilian Symposium on Multimedia and the Web*, pages 335–338. ACM.