



UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO  
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA  
ESCOLA DE INFORMÁTICA APLICADA

SOBRE O TWITTER COMO REDE DE DISCUSSÕES POLÍTICAS EM TEMPO REAL  
NO BRASIL

BIANCA JOAQUIM ALBUQUERQUE DE MELO

**Orientadora:**  
Prof. Vânia Maria Félix Dias, Dsc.

Rio de Janeiro, RJ - Brasil  
Janeiro de 2014

Sobre o Twitter como Rede de Discussões Políticas em Tempo Real no Brasil

Bianca Joaquim Albuquerque de Melo

Projeto de Graduação apresentado à Escola de Informática Aplicada da Universidade Federal do Estado do Rio de Janeiro (UNIRIO) para obtenção do título de Bacharel em Sistemas de Informação.

Aprovado por:

---

Prof. Vânia Maria Félix Dias, Dsc.  
Orientadora

---

Prof. Adriana Cesário de Faria Alvim, Dsc.  
Universidade Federal do Estado do Rio de Janeiro

---

Prof. Mariano Pimentel, Dsc.  
Universidade Federal do Estado do Rio de Janeiro

---

Prof. Adriana Pimenta de Figueiredo, Dsc.  
Universidade Federal do Estado do Rio de Janeiro

Rio de Janeiro, RJ - Brasil

Janeiro de 2014

# **Dedicatória**

Dedico este trabalho aos meus pais, Alzira e Reinaldo, pelo amor e apoio incondicionais e por sempre acreditarem em mim.

# Agradecimentos

Esta monografia é o fruto de uma jornada de quatro anos e meio de graduação, e sua realização não seria possível sem as pessoas e entidades abaixo, às quais gostaria de prestar meus sinceros agradecimentos.

Primeiramente, gostaria de agradecer a Deus por me guiar em cada decisão.

À minha orientadora, Professora Vânia Dias, por toda a dedicação, atenção e paciência ao longo da realização do projeto.

Aos membros da banca examinadora, Professora Adriana Alvim, Professor Mariano Pimentel e Professora Adriana Pimenta, por aceitarem o convite e pela disposição em avaliar e contribuir com este trabalho.

A todo o corpo docente do CCET-UNIRIO, pela contribuição à minha formação acadêmica e por transmitir aprendizados que levarei para toda a vida.

Aos amigos Filipe, André, Jean, Roberta e Amanda, presentes desde o início desta jornada em momentos de estudo, confraternização e de apoio mútuo, principalmente durante os últimos meses.

Aos amigos Danielle, Pedro, Luísa, Jonatas, Wagner, Fernando, Renan, Vilson, Ricardo, Lucas e Makis, que torceram por mim e me apoiaram independente da distância ou da pouca afinidade com o projeto.

À CAPES e ao programa Ciência Sem Fronteiras pela bolsa de estudos no exterior, sem a qual eu não teria a oportunidade de conhecer a Professora Cynthia Hood, que me apresentou ao NodeXL, o Professor Charles Bauer, que me passou uma parcela de toda a sua experiência docente e a professora Madeleine England, que recomendou o livro *Adventures of an IT Leader*, responsável por um grande impacto em minha visão sobre ciência, tecnologia e gestão.

E, por fim, a todos que de algum modo me apoiaram e torceram por mim ao longo desta jornada.

*"Well, you said you think you know some things. What you mean is, you've constructed simplified representations of how those things work. But don't confuse yourself by thinking your simplified mental constructions are realistic, or worse yet, true. [...] You have to judge them by some criteria other than realism. Nothing useful is real. If it's complicated enough to be realistic, it's too complicated to be useful. That's why we build models. Representations. When we say we know things, we just mean we have mental models of those things that we like. Often we like them because they've been useful. But let's not confuse having a useful model with actual knowing."*

*(AUSTIN, Robert D.; NOLAN, Richard L.; O'DONNEL, Shannon. *The Adventures of an IT Leader*. Harvard Business Press, 2009.)*

# Resumo

Análise de redes sociais é uma área relativamente nova que engloba conceitos de diversas outras áreas. Neste trabalho, serão utilizados conceitos oriundos da teoria dos grafos para auxiliar na identificação de vértices-chave das amostras previamente coletadas. Para a realização deste trabalho, coletou-se mais de 16.000 postagens no Twitter ao longo de três dias, com a finalidade de monitorar e analisar o comportamento dos usuários que postaram sobre o voto de desempate do Ministro Celso de Mello, do STF, a favor dos embargos infringentes do processo da Ação Penal 470, popularmente conhecido como 'mensalão'. As postagens encontram-se distribuídas entre três amostras, dentre as quais duas foram selecionadas para uma análise mais aprofundada. Neste trabalho, são analisados quatro grafos, sendo eles duas amostras e seus conjuntos de interesse, que são subgrafos das duas amostras. Os conjuntos de interesse foram montados levando em consideração apenas postagens contendo URLs. As análises realizadas com esses grafos mostram tendências que envolvem não apenas as relações entre os valores das métricas, mas também sobre o comportamento dos usuários que se dispõem a postar no Twitter comentários e notícias envolvendo os acontecimentos do cenário político brasileiro.

**Palavras-chave:** Twitter, Política Brasileira, Análise de Redes Sociais

# Abstract

Social network analysis is a relatively new field of study, that embraces many concepts from other fields. In this work, we use concepts from graph theory to identify key actors in each one of the samples previously collected. Over 16,000 posts on Twitter were collected during three days, in order to monitor and analyze the online behavior of the users who were posting about the casting vote held by Celso de Mello, minister of the Supreme Federal Court, concerning the prosecution 470, best known by its sobriquet 'mensalão', which is a bribery scheme lead by Brazilian politicians. The casting vote concerned the motion for reconsideration, in which the defendants would be allowed a new trial and the possibility of a reduced sentence. The Twitter posts collected are distributed between three samples, among which two were chosen for further analysis. We analyze four graphs, two of them being the samples and the other two being the groups of interest. The groups of interest are both subgraphs of the samples, that take in consideration only posts containing URLs. The results from the analysis show trends that not only concern numerical relations between metrics, but also the behavior of the users who are willing to post on Twitter comments and news about the Brazilian political landscape.

**Keywords:** Twitter, Brazilian Politics, Social Network Analysis

# Sumário

## **Lista de Figuras**

<b>1</b>	<b>Introdução</b>	p. 12
<b>2</b>	<b>Fundamentos teóricos</b>	p. 14
2.1	Definições preliminares . . . . .	p. 14
2.2	Métricas específicas de vértice . . . . .	p. 18
2.2.1	Coeficiente de Agrupamento . . . . .	p. 18
2.2.2	Centralidade de intermediação . . . . .	p. 19
2.2.3	Centralidade de proximidade . . . . .	p. 20
2.2.4	Centralidade de autovetor . . . . .	p. 21
2.2.5	PageRank . . . . .	p. 22
2.3	Métricas de grafo . . . . .	p. 23
2.3.1	Comprimento do Caminho Característico . . . . .	p. 23
2.3.2	Coeficiente de Agrupamento . . . . .	p. 24
<b>3</b>	<b>Coleta</b>	p. 26
3.1	Instrumentos utilizados . . . . .	p. 26
3.1.1	Twitter . . . . .	p. 26
3.1.2	NodeXL . . . . .	p. 28
3.2	Critérios de coleta e organização dos dados . . . . .	p. 33

3.3	Descrição dos grafos . . . . .	p. 35
3.3.1	Grafo STF . . . . .	p. 36
3.3.2	Grafo AP470 . . . . .	p. 36
3.3.3	Grafo #CelsoDeMello . . . . .	p. 38
<b>4</b>	<b>Análise</b>	
4.1	Identificação dos vértices-chave . . . . .	p. 41
4.2	Participação dos vértices-chave nos outros grafos . . . . .	p. 43
4.3	Análise comparativa dos grafos . . . . .	p. 44
4.3.1	Grafo e Conjunto de Interesse STF . . . . .	p. 44
4.3.2	Grafo e Conjunto de Interesse AP470 . . . . .	p. 46
<b>5</b>	<b>Conclusão</b>	
<b>6</b>	<b>Referências</b>	
<b>7</b>	<b>Apêndice</b>	

# **Lista de Figuras**

2.1	Grafo $G_1$	p. 15
2.2	Grafo $G_2$	p. 16
2.3	Grafo $G_3$	p. 17
2.4	Grafo $G_4$	p. 17
2.5	Grafo $G_5$	p. 18
2.6	Grafo $G_6$	p. 19
2.7	Grafo $G_7$	p. 20
2.8	Grafo $G_8$	p. 21
2.9	Grafo $G_9$	p. 24
2.10	Grafo $G_{10}$	p. 25
3.1	Tela inicial do <i>NodeXL</i>	p. 29
3.2	Opções do recurso <i>Autofill Columns</i> para arestas, vértices e grupos	p. 30
3.3	Opções de importação de dados do <i>NodeXL</i>	p. 31
3.4	Tela contendo as opções para a importação da rede de busca por um termo no <i>Twitter</i>	p. 31
3.5	Tela com as opções de métricas calculadas pelo <i>NodeXL</i>	p. 32
3.6	Processo de obtenção das amostras	p. 35
3.7	a) Grafo STF. b) Conjunto de interesse STF destacado em vermelho.	p. 36
3.8	Conjunto de interesse STF	p. 37
3.9	a) Grafo AP470. b) Conjunto de interesse AP470 destacado em vermelho.	p. 37

3.10 Conjunto de interesse AP470 . . . . .	p. 38
3.11 Grafo #CelsoDeMello . . . . .	p. 38
3.12 Grafo desconexo, contendo apenas as interações do grafo #CelsoDeMello que continham endereços da web nas postagens . . . . .	p. 39

# 1. Introdução

Ao longo dos últimos dez anos, os sites de relacionamento se tornaram muito populares e rapidamente atraíram um grande público. Tais sites impactaram drasticamente a forma que as pessoas se comunicam e se mantém informadas sobre o que acontece ao seu redor. Atualmente, sites de relacionamento como *Facebook*, *Twitter* e *LinkedIn* estão entre os mais acessados mundialmente. De acordo com o site [alexa.com](http://alexa.com), o *Facebook* é o segundo site mais acessado em todo o mundo, e *Twitter* e *LinkedIn* ocupam, respectivamente, décima primeira e décima segunda posições<sup>1</sup>.

Devido à grande popularidade desses sites, eles tem sido frequentemente utilizados para pesquisas envolvendo análise das redes sociais formadas por seus usuários. Para fins de análise, as redes sociais podem ser representadas como grafos e nestas representações tende-se a usar os vértices como atores e as arestas como relacionamentos. De acordo com Serrat (2010), a análise de redes sociais tem o objetivo de proporcionar maior entendimento acerca das redes sociais e de seus participantes, focando nos atores e em seus relacionamentos em um contexto social específico.

Muitos desses estudos envolvem a caracterização do perfil dos usuários e das atualizações postadas pelos mesmos, como a pesquisa realizada por Krishnamurty, Gill e Arlitt (2008) sobre o *Twitter*. Cha et al (2010) desenvolveram uma metodologia para comparar a influência de usuários no *Twitter* e para melhor compreender as dinâmicas da influência ao levar em consideração tópicos e momentos distintos. Recentemente, Backstrom e Kleinberg (2013) realizaram um estudo que analisou as conexões de amigos em comum em casais no *Facebook* e desenvolveram uma métrica chamada *dispersão*, que determina se os amigos em comum do casal pertencem a esferas diferentes de amigos e que pode inferir a probabilidade de um casal se separar<sup>2</sup>.

Foi observado no estudo de Tumasjan et al (2010), realizado a partir de coletas de postagens no *Twitter* durante o período próximo às eleições de 2009 na Alemanha, que muitos usuários

---

<sup>1</sup>Disponível em: <http://alexa.com/topsites>. Acesso em 7 de dezembro de 2013.

<sup>2</sup>Disponível em: <http://migre.me/h2hrq>. Acesso em 15 de dezembro de 2013.

do *Twitter* utilizam o site para falar sobre política. O mesmo estudo também aponta que em períodos próximos às eleições, as postagens sobre política tendem a ser mais numerosas e que o *Twitter* não é utilizado apenas para expressar opiniões políticas, mas também para discutí-las com outros usuários do site. Este estudo e a bem sucedida estratégia de campanha online de Barack Obama, presidente dos Estados Unidos (Williams e Gulati, 2008), realçam o importante papel do *Twitter* no cenário político internacional.

O objetivo deste trabalho é buscar melhor compreensão de como os usuários brasileiros do *Twitter* utilizam a rede social para comentar em tempo real acontecimentos referentes ao cenário político do país, e os mesmos não estão restritos às eleições. Neste estudo, o evento que serviu de base para as coletas realizadas foi o voto de desempate do Ministro Celso de Mello a favor dos embargos infringentes no processo da Ação Penal 470, melhor conhecido pela alcunha de 'mensalão'. Os pormenores do acontecimento e de sua escolha para a realização deste estudo serão descritos futuramente neste trabalho.

Este trabalho está organizado em três grandes etapas, que acompanham o desenvolvimento do mesmo. O capítulo 2 apresenta os fundamentos teóricos utilizados como base para a realização do trabalho. Inicialmente, serão detalhadas as definições preliminares, que são os conceitos básicos de teoria de grafos. Em seguida, serão definidas as principais métricas de vértice e de grafo utilizadas em análise de redes sociais. O capítulo 3 trata sobre os assuntos envolvendo a coleta de dados. Neste capítulo são especificados os instrumentos utilizados e suas principais características e funcionalidades, e também são detalhados os critérios utilizados para a coleta e organização dos dados. No capítulo 4 são descritas as etapas para a realização da análise e os critérios para a identificação dos vértices-chave de cada grafo. Neste capítulo, também comenta-se sobre a participação dos vértices-chave nos outros grafos e são realizadas análises comparativas. No quinto capítulo são feitas as considerações finais sobre o trabalho, que envolvem seus principais resultados, limitações e também possibilidades para trabalhos futuros.

## 2. Fundamentos teóricos

Neste capítulo, serão definidos conceitos fundamentais para a melhor compreensão e realização do trabalho. Primeiramente, serão definidos conceitos sobre grafos e, posteriormente, algumas das métricas mais utilizadas em análise de redes sociais.

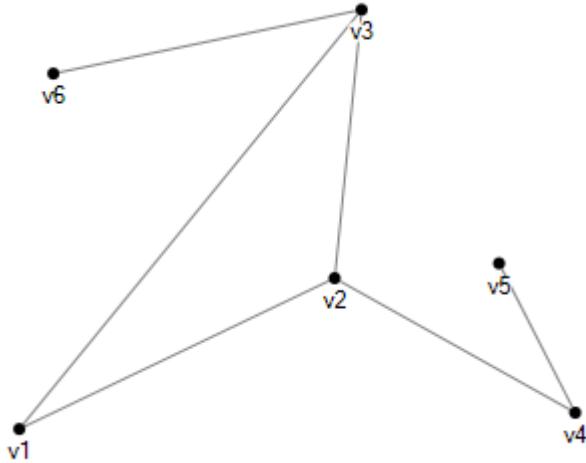
### 2.1 Definições preliminares

Nesta seção são definidos conceitos básicos que serão utilizados ao longo do trabalho.

Um *grafo*  $G = (V, E)$  é um conjunto finito não-vazio  $V$  de  $n$  vértices e um conjunto  $E$  de  $m$  arestas, que são pares não-ordenados de elementos distintos de  $V$ . Dois vértices  $v_1, v_2 \in V$  são *adjacentes* ou vizinhos se existe uma aresta  $e \in E$ , tal que  $e = (v_1, v_2)$ . A *vizinhança* de um vértice  $v \in V$ , denotada por  $\Gamma_v$ , consiste em todos os vértices pertencentes a  $V$  que se conectam a  $v$  por uma aresta. O *grau* de um vértice  $v \in V$ , denotado por  $g(v)$  corresponde ao número de vértices adjacentes a  $v$ . Um *caminho* em um grafo é uma sequência finita de  $k$  vértices  $v_1, v_2, \dots, v_{k-1}, v_k \in V(G)$ ,  $k \geq 1$ , tal que  $(v_j, v_{j+1}) \in E(G)$ ,  $1 \leq j < |k-1|$ ,  $k \leq n$ . A *distância* entre dois vértices  $v_1$  e  $v_k$ , denotada por  $d(v_1, v_k)$  é o comprimento do menor caminho entre  $v_1$  e  $v_k$ .

Por exemplo, seja  $V(G) = \{v_1, v_2, v_3, v_4, v_5, v_6\}$  um grafo com as arestas  $(v_1, v_2)$ ,  $(v_2, v_4)$ ,  $(v_4, v_5)$ ,  $(v_2, v_3)$ ,  $(v_1, v_3)$ ,  $(v_3, v_6) \in E(G)$ . Neste exemplo, o caminho entre  $v_1$  e  $v_5$  é  $v_1, v_2, v_4, v_5$ .

*Tríades*, ou triângulos, ocorrem quando dois vértices adjacentes possuem pelo menos um vizinho em comum. Por exemplo, no grafo  $G_1$ , que pode ser observado na figura 2.1,  $v_1$  é adjacente a  $v_2$  e a  $v_3$ , que também são adjacentes entre si. Portanto, os vértices  $v_1, v_2$  e  $v_3$  formam uma tríade. Um *subgrafo* de  $G$  consiste em um grafo cujos vértices são um subconjunto dos vértices de  $G$ . No grafo  $G_1$  da figura 2.1, o conjunto  $\{v_1, v_3, v_6\}$  é um subgrafo de  $G_1$ .

Figura 2.1: Grafo  $G_1$ 

No grafo  $G_1$ , na figura 2.1, o vértice  $v_2$  possui três vizinhos, sendo eles:  $v_1, v_3$  e  $v_4$ . O grau de um vértice corresponde à quantidade de vizinhos que ele possui, portanto  $g(v_2) = 3$ . Em  $G_1$ , existem dois caminhos possíveis entre  $v_1$  e  $v_6$ . Um desses caminhos é  $v_1, v_3, v_6$  e o outro é  $v_1, v_2, v_3, v_6$ . A distância entre  $v_1$  e  $v_6$  será o comprimento do caminho mais curto entre eles, que é  $v_1, v_3, v_6$ . Portanto,  $d(v_1, v_6) = 2$ .

*Matriz de adjacência* de um grafo é uma forma de representar, através de uma matriz, as conexões entre os vértices do mesmo. Um grafo  $G$  com  $n$  vértices é representado em uma matriz  $n \times n$   $M(G) = m[i, j]$ , em que  $i, j \leq n$ . O valor  $m[i, j]$  informa se os vértices  $v_i$  e  $v_j$  são adjacentes. Se há uma aresta entre  $v_i$  e  $v_j$ ,  $m[i, j] = 1$ . Caso contrário,  $m[i, j] = 0$ .

Pela definição, temos que:

$$m[i, j] = 1, \text{ se } (v_i, v_j) \in E(G)$$

$$m[i, j] = 0, \text{ se } (v_i, v_j) \notin E(G)$$

Por exemplo, o grafo visto na figura 2.1 é representado pela seguinte matriz de adjacência:

$$\begin{pmatrix} & v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \\ v_1 & 0 & 1 & 1 & 0 & 0 & 0 \\ v_2 & 1 & 0 & 1 & 1 & 0 & 0 \\ v_3 & 1 & 1 & 0 & 0 & 0 & 1 \\ v_4 & 0 & 1 & 0 & 0 & 1 & 0 \\ v_5 & 0 & 0 & 0 & 1 & 0 & 0 \\ v_6 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Um grafo  $G = (V, E)$  é *conexo* se existe pelo menos um caminho entre qualquer par de vértices pertencente a  $V(G)$ . Quando não há caminho entre pelo menos um vértice  $v_i \in V(G)$  e qualquer outro vértice do grafo, ele é considerado um grafo *desconexo*. Grafos conexos possuem apenas uma *componente conexa*, que é um subgrafo conexo maximal de  $G$ . Em grafos conexos, a componente conexa coincide com o grafo. Se um grafo possui mais de uma componente conexa, ele é desconexo.

Por exemplo, seja  $G = (V, E)$  um grafo, tal que  $V(G) = \{v_1, v_2, v_3, v_4, v_5, v_6\}$  e  $(v_1, v_2), (v_2, v_3), (v_3, v_4), (v_1, v_3), (v_5, v_6) \in E(G)$ . Neste exemplo,  $G$  é um grafo desconexo que possui duas componentes conexas, uma contendo os vértices  $v_1, v_2, v_3$  e  $v_4$  e as arestas  $(v_1, v_2), (v_2, v_3), (v_3, v_4)$  e outra contendo  $v_5$  e  $v_6$  e a aresta  $(v_5, v_6)$ , como visto na figura 2.2.

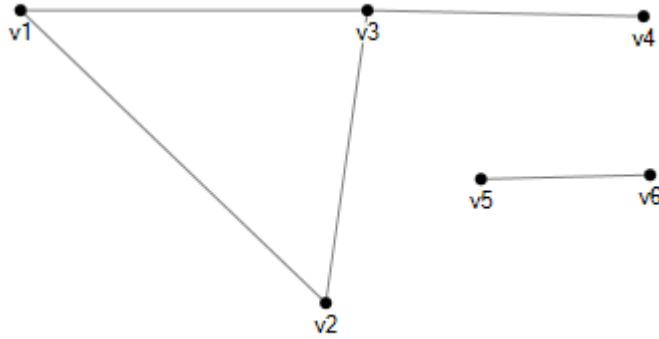


Figura 2.2: Grafo  $G_2$

O grafo  $G_2$ , mostrado na figura 2.2, é um exemplo de grafo desconexo que contém duas componentes conexas. O grafo  $G_1$  (figura 2.1) é um grafo conexo.

Um grafo  $G$  é *não direcionado* se  $v_1, v_2 \in V(G), (v_1, v_2), (v_2, v_1) \in E(G)$  e  $(v_1, v_2) = (v_2, v_1)$ . Caso  $(v_1, v_2) \neq (v_2, v_1)$ , o grafo é considerado *direcionado*, isto é, um grafo cujas arestas pos-

suem direção. Em um grafo direcionado  $G$ , tal que  $V(G) = \{v_1, v_2, v_3, v_4\}$  e  $(v_1, v_2), (v_2, v_3) \in E(G)$ , ao levar em consideração a aresta  $(v_1, v_2)$ ,  $v_1$  é o vértice de origem e  $v_2$  é o vértice de destino. Todavia, ao levar em consideração a aresta  $(v_2, v_3)$ ,  $v_2$  assume a posição de vértice de origem e  $v_3$  de vértice de destino. O *grau de entrada* de um vértice corresponde à quantidade de vezes em um grafo que aquele vértice assumiu a posição de vértice de destino. Já o *grau de saída* de um vértice corresponde ao número de vezes em que aquele vértice foi considerado um vértice de origem.

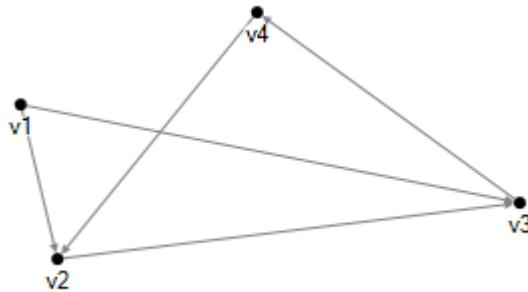


Figura 2.3: Grafo  $G_3$

O grafo  $G_3$ , na figura 2.3, é um exemplo de grafo direcionado. O vértice  $v_1$  possui grau de saída 2 e grau de entrada zero, o grau de entrada de  $v_2$  é 2 e o de saída é 1, assim como  $v_3$ . E o grau de entrada de  $v_4$  corresponde ao mesmo valor do grau de saída, ambos iguais a 1.

O *diâmetro* de um grafo é a maior distância observada entre qualquer par de vértices. Por exemplo, no grafo  $G_4$ , na figura 2.4, o diâmetro do grafo é 3, que corresponde à distância entre  $v_4$  e  $v_5$ , que é a maior distância observada entre todos os pares de vértices do grafo. A *densidade* de um grafo  $G$  é a proporção entre a quantidade de arestas de  $G$  e a quantidade máxima de arestas possíveis entre todos os vértices do grafo.

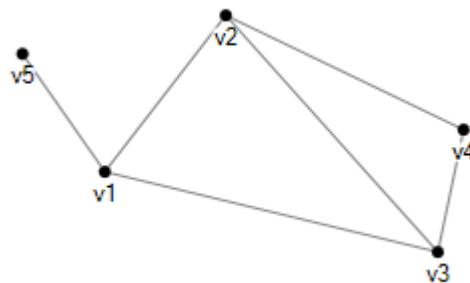


Figura 2.4: Grafo  $G_4$

No exemplo da figura 2.4, a densidade de  $G_4$  é igual a  $\frac{6}{\binom{5}{2}} = \frac{6}{10} = 0,6$ .

*Multigrafos* são grafos que podem possuir múltiplas arestas entre um mesmo par de vértices. Multigrafos também podem possuir *laços*. Um laço é uma aresta que liga um vértice  $v \in V$  a si mesmo. Todos os grafos analisados neste trabalho são multigrafos direcionados e conexos.



Figura 2.5: Grafo  $G_5$

O grafo  $G_5$ , mostrado na figura 2.5, é um exemplo de multigrafo não direcionado, em que o vértice  $v_2$  é adjacente a si mesmo.

A seguir, serão definidas as métricas utilizadas ao longo deste trabalho, que são algumas das mais utilizadas na análise de redes sociais. Algumas métricas são específicas para grafos, e outras de vértices. As fórmulas aqui apresentadas indicam uma das possíveis formas de calcular aproximações dessas métricas. O principal objetivo das definições aqui apresentadas é identificar o que cada métrica representa em análise de redes sociais e quais os parâmetros levados em consideração para o cálculo das mesmas.

## 2.2 Métricas específicas de vértice

Nesta seção, serão definidas as métricas específicas de vértices utilizadas ao longo do trabalho. As métricas de vértice são necessárias para classificar os vértices em um grafo entre si, principalmente as que correspondem a diferentes tipos de centralidade.

### 2.2.1 Coeficiente de Agrupamento

Suzuki e Ribeiro (2007) definem o coeficiente de agrupamento de um vértice  $v$  como a métrica que mede a proporção de vizinhos de  $v$  que também são vizinhos entre si. Quanto maior o coeficiente de agrupamento de um vértice, maior a quantidade de triâdes a que ele pertence. O coeficiente de agrupamento de  $v$ , denotado por  $\gamma(v)$ , é calculado da seguinte forma:

$$\gamma(v) = 1, \text{ se } g_v = 1$$

$$\gamma(v) = \frac{e_{\Gamma v}}{p_{\Gamma v}}, \text{ se } g_v \geq 2$$

Onde  $e_{\Gamma v}$  corresponde à quantidade de arestas entre os vizinhos de  $v$  e  $p_{\Gamma v}$  à quantidade máxima possível de arestas entre os vizinhos de  $v$ , o que corresponde à combinação dois a dois do número de vizinhos de  $v$ . Portanto,  $p_{\Gamma v} = \binom{g_v}{2}$ .

Se  $v$  possui apenas um vizinho, o coeficiente de agrupamento é 1. Quando  $v$  possui mais de um vizinho, o coeficiente de agrupamento é calculado por uma proporção entre o número de arestas incidentes entre os vizinhos de  $v$  e o número máximo de arestas que poderiam incidir entre os vizinhos de  $v$ . O cálculo também poderia ser feito através da proporção entre a quantidade de triângulos contendo  $v$  e a quantidade máxima de triângulos que poderiam conter  $v$  como um de seus vértices. As duas formas de cálculo estão corretas e possuem o mesmo resultado, embora a última seja mais útil para grafos pequenos e de fácil visualização.

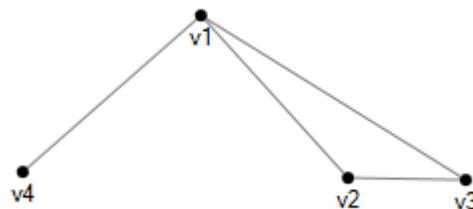


Figura 2.6: Grafo  $G_6$

Para calcular o coeficiente de agrupamento do vértice  $v_1$  do grafo  $v_6$ , na figura 2.6,  $e_{\Gamma v_1} = 1$  e  $p_{\Gamma v_1} = 3$ . Portanto,  $\gamma(v_1) = \frac{1}{3}$ .

## 2.2.2 Centralidade de intermediação

Centralidade de intermediação, ou *betweenness centrality*, é a métrica que define o grau de participação de um vértice nos caminhos mais curtos de um grafo. Brandes e Pich (2006) associam o conceito de centralidade de intermediação à ideia do controle sobre as conexões entre outros pares de vértices no grafo.

Denota-se por  $\sigma(s, t)$  a quantidade de caminhos distintos mais curtos entre  $s$  e  $t$ , e  $\sigma(s, t|v)$  o número de caminhos mais curtos entre  $s$  e  $t$  que contém  $v$  como intermediário.  $v \neq s$  e  $v \neq t$ ,

tal que  $v, s, t \in V(G)$ . A centralidade de intermediação é calculada da seguinte forma:

$$Bet(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}, s \neq v \neq t$$

Para o cálculo da centralidade de intermediação de um vértice  $v$ , é feita a proporção entre a quantidade de caminhos mais curtos que contém  $v$  como intermediário e a quantidade total de caminhos mais curtos entre dois outros vértices quaisquer do grafo. Para fins de transmissão de notícias, altos valores de centralidade de intermediação indicam vértices que podem transmitir notícias de forma mais rápida que os demais, e também vértices com a maior probabilidade de receber certas notícias mais rapidamente, já que intermedeia a maioria dos vértices e possui um fluxo de informação maior que o de outros vértices.

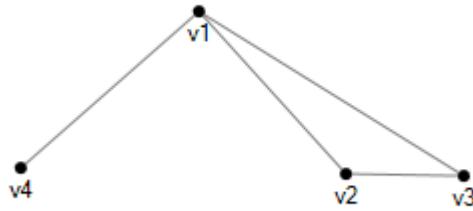


Figura 2.7: Grafo  $G_7$

Para calcular a centralidade de intermediação do vértice  $v_1$  do grafo  $G_7$ , na figura 2.7, é necessário contar a quantidade de caminhos mais curtos entre quaisquer dois vértices de  $G_7$  e também quantos deles contém  $v_1$  como intermediário. Há um total de três caminhos mais curtos entre os vértices diferentes de  $v_1$ , e dois desses caminhos possuem  $v_1$  como intermediário. Portanto, a centralidade de intermediação de  $v_1$  é igual a  $\frac{2}{3}$ , isto é, das três conexões entre os vértices  $v_2, v_3$  e  $v_4$ , duas delas são intermediadas por  $v_1$ .

### 2.2.3 Centralidade de proximidade

A centralidade de proximidade, ou *closeness centrality*, mede a proximidade de um determinado vértice a todos os outros contidos no grafo. Esta medida procura determinar se um vértice é ou não central. Um vértice é central se pode interagir rapidamente com todos os outros (Wasserman e Faust, 1994), isto é, se encontra relativamente próximo a todos os outros.

O cálculo da centralidade de proximidade é feito da seguinte maneira:

$$Clo(v) = \frac{1}{\sum_{t \in V} d(v, t)}$$

Para calcular a centralidade de proximidade, é necessário determinar a distância do vértice  $v$  a todos os outros vértices do grafo e calcular o somatório das distâncias obtidas. Posteriormente, o valor é normalizado para obter valores de centralidade de proximidade entre zero e 1. Quanto maior o valor da centralidade de proximidade do vértice, mais próximo aos outros vértices do grafo ele se encontra.

A centralidade de proximidade leva em consideração as distâncias do vértice escolhido aos demais, isto é, mede a potencial eficiência de  $v$ , mesmo em um cenário de pior caso. Um vértice com alto valor de centralidade de proximidade está, em média, mais próximo a todos os outros vértices do grafo, o que o torna um vértice importante para comunicação eficiente.

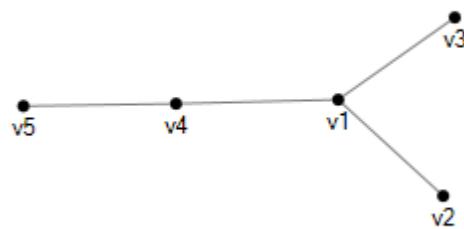


Figura 2.8: Grafo  $G_8$

Para exemplificar o cálculo da centralidade de proximidade, será utilizado o vértice  $v_1$  do grafo  $G_8$ , na figura 2.8. O somatório das distâncias de  $v_1$  aos outros vértices do grafo é 5, portanto o valor da centralidade de proximidade de  $v_1$  é 0,2. A centralidade de proximidade do vértice  $v_4$  é  $\frac{1}{1+1+2+2} = \frac{1}{6}$ , o que significa que  $v_4$  está, de um modo geral, menos próximo dos outros vértices de  $G_8$ .

## 2.2.4 Centralidade de autovetor

A centralidade de autovetor, ou *eigenvector centrality*, é uma métrica que calcula a importância de um vértice em um grafo de acordo com a sua matriz de adjacência (Wang, Scaglione e Thomas, 2010). Dado um grafo  $G = (V, E)$ , sua matriz de adjacência  $A$ , um autovalor  $\lambda$  é um

escalar que satisfaz a seguinte igualdade:

$$\lambda \cdot x = A \cdot x$$

Dizemos que  $x$  é um autovetor associado ao autovalor  $\lambda$ .

A centralidade do vértice  $v$  é definida como a  $v$ -ésima posição do autovetor  $x$  correspondente ao maior autovalor  $\lambda_{max}$ .

$$C_e(v) = x_v = \frac{1}{\lambda_{max}} \sum_{j=1}^n A(v, j)x_j$$

A centralidade de  $v$  é proporcional à soma das centralidades de todos os seus nós vizinhos. A definição escolhe o autovetor correspondente ao maior autovalor  $\lambda_{max}$  para garantir que todos os valores de centralidade sejam positivos.

Os valores não-diagonais que se encontravam na matriz  $Y$  podem ser vistos como a força de conectividade. Tais valores foram eliminados de  $A$  na diagonalização, mas é possível reintegrá-los à matriz de adjacência da seguinte forma:

$$A_Y = -Y + D(Y)$$

Sendo  $D(Y)$  a matriz diagonal obtida a partir da matriz original.

O cálculo da medida de centralidade é dado pela magnitude dos valores do autovetor:

$$C_e Y(V) = ||x_v|| = \left| \left| \frac{1}{\lambda_{max}} \sum_{j=1}^n A_Y(v, j)x_j \right| \right|$$

## 2.2.5 PageRank

A métrica *PageRank* foi desenvolvida com base na ideia de que, em um grafo direcionado, uma aresta que parte do vértice  $v_1$  para o vértice  $v_2$  indica que o autor de  $v_1$  demonstrou interesse em  $v_2$ . Deste modo, se um vértice é muito referenciado por outros, pode-se concluir que há um grande interesse por esse vértice e que ele deve ser considerado importante ou com conteúdo de alta qualidade. Além disso, é esperado que o conteúdo de um vértice importante seja mais significativo do que o conteúdo de um vértice aleatório.

Cho e Roy (2004) definem *PageRank* como a importância do vértice  $v$ , calculada a partir da soma da importância dos vértices que apontam para  $v$ . Se muitos vértices importantes referenciarem  $v$ , o *PageRank* de  $v$  será alto. Esta métrica é frequentemente utilizada em serviços de busca, de modo que as páginas com maior *PageRank* são consideradas potencialmente mais

relevantes como resultados de uma busca. Em análise de redes sociais, o *PageRank* é utilizado para classificar cada um dos vértices de um grafo de acordo com a sua relevância.

Para exemplificar o cálculo do *PageRank* para um vértice, será utilizado o vértice  $v_i$  que é referenciado pelas páginas  $v_1, \dots, v_n$ . Seja  $l_j$  o número de arestas partindo do vértice  $v_j$ . Seja  $d$  o fator de dispersão. O cálculo do *PageRank* de  $v_i$  é dado por:

$$P\text{rank}(v_i) = d + (1 - d)(P\text{rank}(v_1)/l_1 + \dots + P\text{rank}(v_n)/l_n)$$

O cálculo do *PageRank* possui diversas versões, sendo a apresentada acima um modelo mais simplificado, mas que leva em consideração a abstração de um usuário visitando um vértice (página) e com a probabilidade  $d$  de que o próximo vértice a ser visitado será completamente aleatório, pois eventualmente o usuário vai se distrair e acessar outras páginas não interligadas ao conjunto anterior. Dado isso,  $1 - d$  é a probabilidade de o próximo vértice a ser visitado ser um dos  $l_i$  referenciados por  $v_i$ .

As métricas de vértice descritas nesta seção serão utilizadas posteriormente neste trabalho, na etapa de identificação de vértices-chave de cada grafo. A seguir, serão definidas as métricas de grafo, que também serão utilizadas ao longo da análise dos grafos.

## 2.3 Métricas de grafo

Nesta seção, serão definidas as métricas de grafo utilizadas neste trabalho. Enquanto as métricas de vértice são utilizadas para classificar os vértices de um grafo conforme algum critério específico de importância, as métricas de grafo proporcionam informações úteis para a comparação entre dois ou mais grafos.

### 2.3.1 Comprimento do Caminho Característico

O comprimento do caminho característico  $L(G)$  é a mediana das médias das menores distâncias entre cada vértice  $v \in V(G)$  e todos os outros. Seja  $d(u, v)$  a distância entre o vértice fixo  $u$  e cada vértice  $v$  do grafo.

O comprimento do caminho característico de um grafo  $G$  é calculado da seguinte forma:

$$L(G) = \text{mediana}_{u \in V} \left( \frac{\sum_{v \in V(G)} d(u, v)}{n} \right)$$

Para calcular o comprimento do caminho característico, é necessário calcular as menores distâncias entre cada um dos vértices do grafo e todos os outros. Computacionalmente, é necessário executar uma busca por largura para cada vértice para obter tais valores. Posteriormente, é calculada a distância média de cada vértice para todos os outros do grafo. Após este cálculo, tais valores são ordenados para a obtenção da mediana. O valor encontrado para a mediana corresponderá ao valor do comprimento do caminho característico.

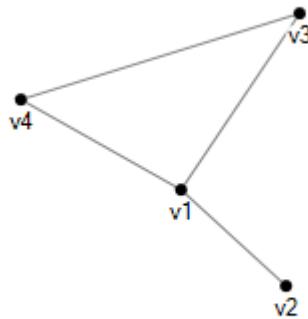


Figura 2.9: Grafo  $G_9$

Para exemplificar o cálculo do comprimento do caminho característico de um grafo, será utilizado o grafo  $G_9$ , visto na figura 2.9. A média das menores distâncias de  $v_1$  é 1, de  $v_2$  é  $\frac{5}{3}$ , de  $v_3$  é  $\frac{4}{3}$  e  $v_2$  é  $\frac{4}{3}$ . Ordenados, temos 1,  $\frac{4}{3}$ ,  $\frac{4}{3}$  e  $\frac{5}{3}$ . Portanto, o comprimento do caminho característico de  $G_9$  é  $\frac{4}{3}$ .

Esta métrica é importante para a obtenção de uma estimativa de eficiência na comunicação de um grafo, o que justifica o uso da mediana e não de um cálculo de média das médias, já que a mediana é uma medida que retrata de forma mais fiel comportamentos numéricos.

### 2.3.2 Coeficiente de Agrupamento

O coeficiente de agrupamento  $C(G)$  do grafo é uma medida do valor relativo de triângulos existentes no grafo e é determinado a partir da média entre os coeficientes de agrupamento de

todos os seus vértices.

$$C(G) = \frac{\sum_{v \in V} \gamma(v)}{n}$$

Para calcular o coeficiente de agrupamento do grafo, é necessário calcular o coeficiente de agrupamento de cada um dos vértices do grafo. Depois disso, é calculada a média da soma desses valores para a obtenção de  $C(G)$ . O cálculo do coeficiente de agrupamento de um vértice foi mostrado na seção 1.2.1.

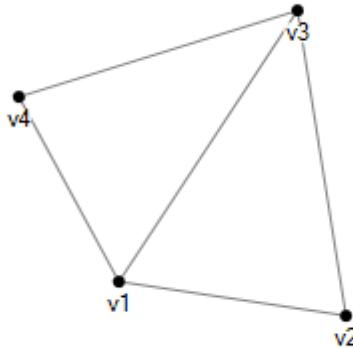


Figura 2.10: Grafo  $G_{10}$

Na figura 2.10, vê-se o grafo  $G_{10}$ , que contém 4 vértices e 5 arestas. Temos  $\gamma(v_1) = \gamma(v_3) = \frac{2}{3}$  e  $\gamma(v_2) = \gamma(v_4) = 1$ . Portanto:

$$C(G_{10}) = \frac{2 + \frac{4}{3}}{4} = \frac{\frac{10}{3}}{4} = \frac{5}{6}$$

Assim como ocorre no cálculo do coeficiente de agrupamento do vértice, um cálculo alternativo para o coeficiente de agrupamento do grafo é calcular a proporção entre a quantidade de tríades existentes no grafo e a quantidade máxima de tríades possíveis. Quanto maior o coeficiente de agrupamento de um grafo, mais curtos são os caminhos entre seus vértices.

Ao longo deste capítulo, foram definidos os conceitos de teoria de grafos e de métricas de grafos e vértices que serão utilizadas para a análise dos grafos obtidos na coleta de dados, cujos procedimentos serão tratados com maior detalhe no capítulo a seguir.

## 3. Coleta

Este capítulo trata dos principais assuntos relacionados à coleta de dados para a realização do trabalho, como a rede social escolhida para a extração dos dados, a ferramenta utilizada para a coleta e a análise e também os critérios utilizados para a coleta e organização dos dados.

### 3.1 Instrumentos utilizados

O objetivo desta seção é identificar os instrumentos utilizados para a coleta dos dados neste trabalho, além de detalhar suas principais características e também justificar a utilização dos mesmos. Os dados foram extraídos da rede social *Twitter*, utilizando o *NodeXL*, template para Microsoft Excel.

#### 3.1.1 Twitter

*Twitter.com* é uma rede social online utilizada por milhões de pessoas em todo o mundo para se manterem em contato com amigos, familiares e colegas de trabalho através de computadores e dispositivos móveis (HUBERMAN, ROMERO e WU, 2008).

As principais utilizações do *Twitter* envolvem transmissão de notícias, cobertura em tempo real de eventos, campanhas políticas, publicidade e organização de protestos e *flash mobs*<sup>1</sup>. Os usuários comentam, principalmente, acontecimentos recentes. O *Twitter* é amplamente utilizado por empresas de diversas áreas, não apenas como mais um canal de interação com clientes, mas também como um canal de notícias e atualizações durante acontecimentos em todo o mundo, como desastres naturais e outras emergências (JANSEN et al, 2009).

A não reciprocidade é uma das características mais marcantes do sistema, que permite que usuários estabeleçam comunicação entre si mesmo se não se seguirem, ou mesmo que apenas

---

<sup>1</sup>”Flash mob, por definição, é uma súbita mobilização coletiva, em espaços públicos físicos, organizada através da internet ou outras redes de comunicação digital.” Disponível em: <http://migre.me/gL8sZ>. Acesso em 22 de novembro de 2013.

um dos envolvidos na conversa siga o outro, e esta característica fez com que o *Twitter* fosse amplamente utilizado por grandes empresas para estabelecer contato direto com seus consumidores.

De acordo com Cassaes e Garcia (2011), o *Twitter* pode ser utilizado como ferramenta que potencializa os fluxos de notícias e quebra o paradigma dos monopólios da informação, o que mostra o potencial do *Twitter* na colaboração entre os usuários. Com o advento do *Twitter*, o fluxo de notícias deixou de ser unidirecional ou regular. O benefício que isso traz para os usuários é o acesso a diversas notícias de seu interesse, oriundas de diversas localizações, quase em tempo real. O usuário do *Twitter* não precisa ser jornalista para produzir informação relevante para seus seguidores.

Jaramillo (2010) ressalta que o *Twitter* está em constante reinvenção, acrescentando recursos e modificando diversos aspectos do sistema para que seus usuários obtenham a melhor experiência. Além disso, também afirma que cada usuário possui sua própria percepção sobre o que seja o *Twitter*, pois o sistema pode ser utilizado para diversas finalidades.

Burge e Comm (2009) frisam a importância do *Twitter* na internet contemporânea, por ser um sistema que desencadeou reações que outros sistemas até então não haviam despertado em seus usuários, criando novas tendências para a comunicação digital. Berti (2011) analisou diversas contas de *Twitter* de sites de notícias do estado do Piauí e constatou que, apesar de reverberarem conteúdos de seus respectivos portais, muitas vezes surgiam conteúdos oriundos do público do jornal, que ao mesmo tempo em que consumia as notícias, retroalimentava o sistema com informações adicionais sobre acontecimentos específicos. Esse contra-fluxo informacional, tal qual especificado pelo autor supracitado, pode ser observado constantemente no *Twitter*, independente do contexto da localização de usuários, e visto como uma retroalimentação, tornando tênue a fronteira entre quem produz e quem consome conteúdo.

O *Twitter* possui alguns termos e conceitos próprios, cujas definições nem sempre correspondem aos significados tradicionalmente conhecidos. Com o passar do tempo, também foram criados alguns neologismos para descrever certos comportamentos na rede. Os termos mais utilizados no *Twitter* e seus respectivos significados serão descritos a seguir, dado que os mesmos também serão utilizados extensivamente ao longo do trabalho. *Tweet* é uma postagem no *Twitter*, e todos os *tweets* devem possuir no máximo 140 caracteres. *Hashtag* é uma forma de indexar palavras-chave em uma postagem. As *hashtags* são palavras precedidas pelo símbolo # e são clicáveis, podendo redirecionar o usuário para uma página que contenha outros *tweets* com o mesmo termo. *Retweet* é o ato de postar novamente algo que já foi postado por outro usuário, se valendo de suas palavras exatas. Os *retweets* possibilitam que as mensagens

postadas por um usuário atinjam um público além de seu conjunto de seguidores. O formato de um *retweet* é RT @usuario mensagem.

*Seguidores* ou *followers* são um grupo de usuários que seguem outros usuários no *Twitter*. Seguir significa assinar o *feed* de atualizações do usuário. A *timeline* é a linha do tempo em que constam as postagens de todos os perfis que um usuário segue. As postagens aparecem em ordem cronológica inversa, com os *tweets* mais recentes no topo.

A reciprocidade das conexões entre usuários não é obrigatória no *Twitter*, portanto as arestas que representam tais conexões são sempre direcionadas. O *Twitter* possui algumas métricas individuais cuja obtenção é direta, como o número de seguidores e a quantidade de usuários que ele segue. Apesar de as definições dessas métricas serem complementares, é importante ressaltar que elas são independentes uma da outra e que não há nenhuma relação numérica entre as mesmas.

Cada usuário do *Twitter* tem a possibilidade de seguir usuários e de ter seguidores. Tais papéis não são excludentes e é possível seguir e ser seguido, sem limites quantitativos. Vale ressaltar que não há a obrigação da reciprocidade nos relacionamentos no *Twitter*, portanto é possível um usuário seguir quantos perfis quiser e ter poucos seguidores. Com o limite da quantidade de caracteres, a forma mais eficaz de disseminar conteúdo utilizando o *Twitter* é através da postagem de URLs apontando para outros sites. Atualmente, todos os *links* postados no *Twitter* são automaticamente encurtados pelo próprio encurtador do site, o t.co.

As principais características que influenciaram na escolha desta rede social foram as postagens mais curtas e a não obrigatoriedade para com a reciprocidade dos relacionamentos. Outros fatores também foram ponderados, como a facilidade para coletar postagens e interações contendo termos específicos e a simplicidade de representar as interações através de grafos. A seguir, serão descritas as principais funcionalidades do *NodeXL*, a ferramenta escolhida para a coleta, organização e análise dos dados utilizados neste trabalho.

### 3.1.2 NodeXL

O *NodeXL* é uma ferramenta *open-source* gratuita, desenvolvida como um template para o programa Microsoft Excel, que não apenas permite a entrada manual de valores em suas planilhas, mas também possui extensões que possibilitam a coleta de dados de redes de usuários diretamente de sites como *Twitter*, *Facebook*, *Flickr* e *YouTube*. Neste trabalho, foi utilizada a versão 1.0.1.245, lançada em 19 de junho de 2013<sup>2</sup>.

---

<sup>2</sup>Disponível em: <http://nodexl.codeplex.com/>. Acesso em 19 de novembro de 2013.

De acordo com Hansen, Shneiderman e Smith (2011), o *NodeXL* foi desenvolvido principalmente para facilitar o aprendizado de conceitos e métodos de redes sociais, utilizando a visualização como um componente-chave. Além disso, o template faz uso de diversas planilhas que contém campos para armazenar todas as informações necessárias para representar um grafo.

Há uma planilha apenas para os vértices e as arestas são listadas em outra planilha dentro do mesmo documento. Outras planilhas contém métricas do grafo, que são calculadas pela ferramenta, e também informações sobre grupos e componentes conexas. As funcionalidades de visualização permitem alterar propriedades visuais de vértices e arestas, como cor, tamanho, espessura e forma.

A principal vantagem em se usar o *NodeXL* é ter detalhes do grafo organizado em planilhas do Microsoft Excel, o que permite a realização de análises estatísticas mais elaboradas sem a necessidade de exportar os dados para outra ferramenta. Todas as funcionalidades do Excel podem ser aplicadas às planilhas. O *NodeXL* funciona em ambos Microsoft Excel 2007 e 2010, mas não está disponível em versões para Mac. O *NodeXL* é uma ferramenta versátil e repleta de funcionalidades que possibilitam a realização de todas as etapas da análise de uma rede social, sem a necessidade de importar ou exportar dados para outras ferramentas.

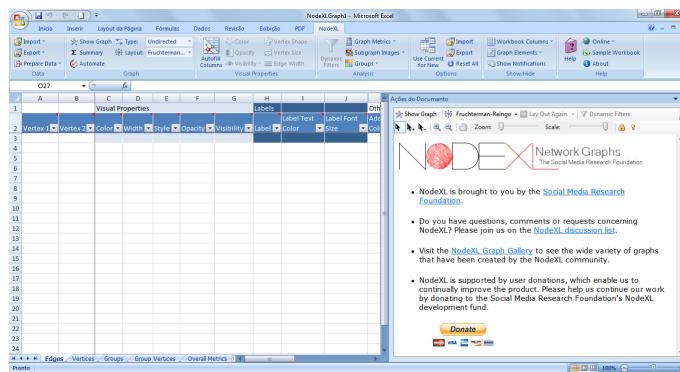


Figura 3.1: Tela inicial do *NodeXL*

Dentre as funcionalidades básicas do *NodeXL*, se destacam a inserção manual de vértices e arestas de um grafo nas células das planilhas. A qualquer momento é possível gerar a imagem do grafo correspondente aos dados que se encontram na planilha, e também definir se o grafo será direcionado ou não e qual o algoritmo utilizado para posicionar os vértices e arestas na tela. Caso seja necessário, os vértices podem ser ajustados manualmente pelo usuário para que assumam posições mais convenientes para a visualização do grafo ou de um de seus conjuntos de vértices.

Além disso, o *NodeXL* possui diversos algoritmos para agrupar os vértices dos grafos con-

forme os critérios desejados. Por exemplo, é possível agrupar os vértices por componente conexa, por atributos do vértice, como grau e outras métricas, por motivo e também por grupo. Para agrupamentos por grupo, pode-se escolher entre os algoritmos de Causet-Newman-Moore, Wakita-Tsurumi e Girvan-Newman, sendo o último recomendado apenas para grafos menores por ter maior tempo de execução.

O *NodeXL* é repleto de propriedades visuais, responsáveis por alterar a aparência de vértices e arestas e facilitar o destaque visual de elementos importantes no grafo. Além de configurar manualmente atributos como cor, tamanho, forma e espessura de vértices e arestas, também é possível alterar tais atributos automaticamente utilizando o *Autofill Columns*, recurso de preenchimento automático de colunas que modifica as propriedades visuais dos grafos de acordo com os critérios determinados pelo usuário. É possível, por exemplo, utilizar a funcionalidade *Autofill Columns* para fazer com que o tamanho de cada um dos vértices de um grafo seja diretamente proporcional a seus respectivos graus. A automatização torna essa tarefa substancialmente mais rápida que quando feita manualmente, além de possibilitar a riqueza de detalhes e propriedades visuais mesmo em grafos muito grandes.

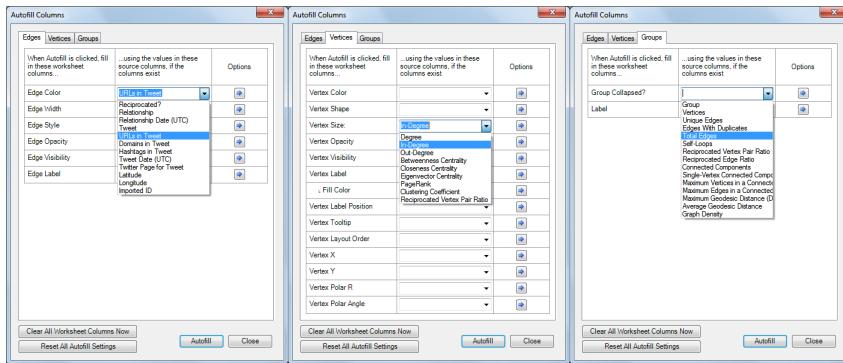


Figura 3.2: Opcões do recurso *Autofill Columns* para arestas, vértices e grupos

Além de alterar atributos visuais de elementos do grafo, o *NodeXL* possibilita ocultar vértices ou arestas com características específicas ou que correspondem a algum determinado critério quando se utiliza o modo filtro. Também é possível fazer isso manualmente para uma quantidade pequena de vértices ou ainda mesmo para ocultar temporariamente vértices isolados sem utilizar nenhum critério específico. Neste trabalho, as análises foram feitas em grafos obtidos a partir da aplicação de filtros. O processo utilizado para a filtragem dos grafos será descrito no próximo capítulo, juntamente com a descrição da coleta dos dados.

Outra importante funcionalidade do *NodeXL* é a importação de redes, que possibilita a coleta de redes diretamente de sites populares como *Twitter*, *Facebook*, *YouTube* e *Flickr* para análise. Além disso, o *NodeXL* também pode importar redes obtidas das contas de *e-mail* sincronizadas com o Microsoft *Outlook*, além de possibilitar também a importação de redes

coletadas com outras ferramentas.

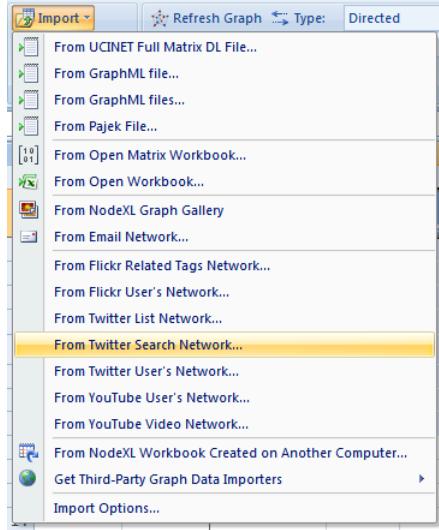


Figura 3.3: Opções de importação de dados do *NodeXL*

Neste trabalho, foram importados dados do *Twitter* e a rede foi obtida a partir da busca por três termos diferentes. A ferramenta não é *case-sensitive* e não reconhece acentuação. Para fazer uma coleta usando o termo 'presidência', por exemplo, deve-se buscar por 'presidencia', ou o conjunto retornado será referente ao termo 'presid'. Para coletar apenas as ocorrências de uma *hashtag*, é necessário incluir o símbolo # antes do termo. A não inserção do símbolo torna a coleta menos restritiva. Utilizando novamente o exemplo do termo 'presidência', a busca por '#presidencia' irá incluir apenas *hashtags*, enquanto a busca por 'presidencia' incluirá todas as ocorrências do termo.

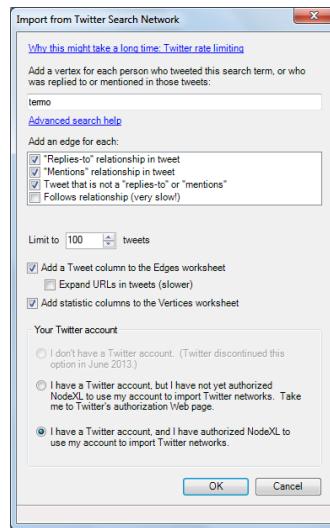


Figura 3.4: Tela contendo as opções para a importação da rede de busca por um termo no *Twitter*

O *NodeXL* também realiza o cálculo de métricas de vértices, grafos e de grupos. O *NodeXL* calcula tanto métricas gerais do grafo como também métricas específicas de cada vértice.

Alguns exemplos das métricas de grafo calculadas pelo *NodeXL* são quantidade de vértices, de arestas com e sem duplicatas, além do número total, taxa de vértices e arestas reciprocados, componentes conexas, número máximo de vértices e arestas em uma componente conexa, comprimento do caminho característico, diâmetro do grafo, diâmetro médio e densidade.

Quanto às métricas específicas de vértices, o *NodeXL* calcula o grau para grafos não direcionados e o grau de entrada e saída para grafos direcionados. Também são calculadas métricas de centralidade, como as de proximidade, intermediação e autovetor. Além dessas, são calculados o coeficiente de agrupamento e o PageRank de cada vértice, entre outras métricas. As métricas e funcionalidades previamente listadas são suficientes para a realização deste trabalho.

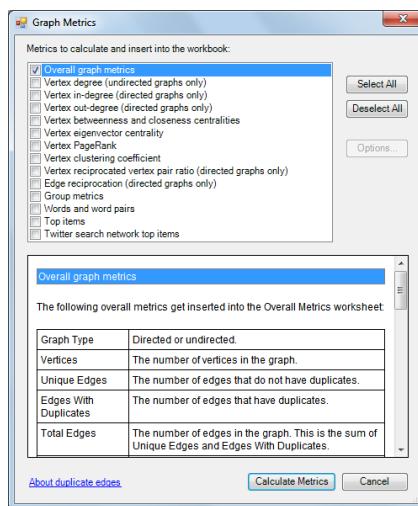


Figura 3.5: Tela com as opções de métricas calculadas pelo *NodeXL*

A escolha do *NodeXL* para a realização deste trabalho se deu principalmente pela versatilidade da ferramenta e pela sua integração com o Microsoft Excel, que possibilita a aplicação de filtros, ordenação e outros recursos úteis tanto para análise quanto para a organização dos dados nas planilhas. Apesar de tantos recursos, o *NodeXL* é uma ferramenta de utilização simples.

Após a decisão da rede social de onde os dados seriam extraídos e da ferramenta de coleta e análise dos dados, foram definidos os critérios a serem utilizados na coleta de dados. Tais critérios envolveram, inicialmente, a escolha de um acontecimento no cenário político com um potencial de alta expressividade na internet, principalmente na rede social desejada, o estabelecimento do período de coleta e dos termos a serem utilizados para a obtenção das redes.

## 3.2 Critérios de coleta e organização dos dados

A coleta foi realizada no *Twitter* durante os dias 17, 18 e 19 de setembro de 2013. Foram feitas diversas coletas para cada termo pesquisado, de forma a obter a maior quantidade possível de postagens em cada um dos conjuntos.

Os critérios da coleta foram definidos com base em acontecimentos recentes no cenário político nacional. No dia 18 de setembro de 2013, o Ministro Celso de Mello, decano do Supremo Tribunal Federal, deu seu voto de desempate a favor da aceitação dos embargos infringentes no processo da Ação Penal 470<sup>3,4</sup>, também conhecido como 'mensalão' ou AP 470.

Antes desta votação, o Plenário da Suprema Corte se encontrava dividido, com cinco votos a favor e cinco votos contra tal decisão. Com a aceitação dos embargos infringentes, 12 réus na Ação Penal 470 têm a possibilidade de entrarem com recurso de condenações pelos crimes de formação de quadrilha e lavagem de dinheiro<sup>5</sup>. Após a revisão de recursos, os réus podem ter suas penas reduzidas ou até mesmo serem absolvidos<sup>6</sup>. Apenas os réus que tiveram pelo menos quatro votos no sentido da absolvição podem entrar com o recurso, portanto todos os processos envolvendo réus que não têm o direito de apresentar embargos infringentes não serão mudados<sup>7,8</sup>.

Tais acontecimentos causaram grande comoção na internet, principalmente em redes sociais, desde o dia em que foi realizada a primeira votação, que terminou empatada<sup>9,10</sup>. Até o momento da votação, o Ministro sofreu grande pressão nas redes sociais, tanto para votar contra quanto a favor da aceitação dos embargos infringentes<sup>11</sup>. Tal tendência foi confirmada pelas postagens nos datasets coletados, datadas até o início da tarde do dia 18 de setembro.

Após encerrado o período de coleta, foram obtidos três datasets preliminares. Cada um desses datasets preliminares representa o resultado bruto da coleta para cada um dos termos pesquisados, sem nenhum tipo de filtro. Como tais datasets continham algumas inconsistências, como spam e postagens duplicadas, foi realizada uma filtragem com o objetivo de eliminar tais ocorrências e evitar resultados inflados. Em um primeiro momento, para facilitar a compreensão do processo, os conjuntos de dados coletados serão chamados de datasets. Os grafos serão

<sup>3</sup>Disponível em: <http://migre.me/gFpQX>. Acesso em 14 de novembro de 2013.

<sup>4</sup>Disponível em: <http://migre.me/gFq0j>. Acesso em 14 de novembro de 2013.

<sup>5</sup>Disponível em: <http://migre.me/gFqbd>. Acesso em 14 de novembro de 2013.

<sup>6</sup>Disponível em: <http://migre.me/gFqcK>. Acesso em 14 de novembro de 2013.

<sup>7</sup>Vide nota de rodapé 5

<sup>8</sup>Disponível em: <http://migre.me/gFqdA>. Acesso em 14 de novembro de 2013.

<sup>9</sup>Disponível em: <http://migre.me/gFqe7>. Acesso em 14 de novembro de 2013.

<sup>10</sup>Disponível em: <http://migre.me/gFqf0>. Acesso em 14 de novembro de 2013.

<sup>11</sup>Disponível em: <http://migre.me/gFqeX>. Acesso em 14 de novembro de 2013.

gerados a partir dos datasets selecionados e, após isso, o trabalho fará referência apenas aos grafos.

É importante destacar a diferença entre postagens duplicadas e arestas duplicadas. Arestas duplicadas são arestas que ligam o mesmo par de vértices na mesma direção mais de uma vez em momentos diferentes. Isso significa que cada aresta carrega um conteúdo diferente dos anteriores. Arestas duplicadas mostram uma maior força de interação entre dois vértices, portanto devem ser mantidas.

Por outro lado, as postagens duplicadas consistem na repetição de uma mesma aresta, com as mesmas propriedades e mesmo conteúdo, na planilha de arestas. Durante a realização das coletas, há uma grande chance de aparecerem repetições de uma mesma postagem, com a mesma mensagem sendo postada no mesmo horário envolvendo os mesmos vértices de origem e destino. Postagens duplicadas são meras repetições que não agregam nada de novo ao dataset, além de impactarem o cálculo das métricas tanto do vértice quanto do grafo, já que a ocorrência das duplicatas não seguiam nenhum padrão específico.

Os datasets filtrados consistem na maior componente conexa obtida após a filtragem dos datasets preliminares. Pelas razões supracitadas, os datasets preliminares possuem muitas inconsistências para serem utilizados neste trabalho. Tais inconsistências não estão presentes nos datasets filtrados, que servirão de base para a obtenção dos **conjuntos de interesse**.

Para os conjuntos de interesse, foi feita mais uma filtragem nos datasets filtrados, de modo a manter apenas as arestas que representavam postagens contendo algum endereço da web (URL) e os vértices envolvidos em tais ligações. Uma vez identificados em cada um dos datasets filtrados, foi obtida a maior componente conexa de cada um destes grafos, que serão chamados de conjuntos de interesse. Em um ambiente de mensagens curtas como o *Twitter*, a propagação de notícias com maiores detalhes é facilitada com o uso de URLs. Principalmente neste trabalho, que tem como ponto de partida um acontecimento de cunho político, a presença de URLs em postagens é um indicativo do grau de envolvimento dos usuários com a discussão do ocorrido. Esse grupo de usuários com maior predisposição a engajar em discussões mais substanciais sobre os termos pesquisados é de grande interesse para uma análise contextual das postagens coletadas.

### 3.3 Descrição dos grafos

Este trabalho procura compreender como a audiência brasileira do *Twitter* utiliza a rede para acompanhar, comentar e discutir acontecimentos e notícias sobre o cenário político brasileiro, mais especificamente sobre os acontecimentos do dia 18 de setembro de 2013, que incluem o voto do Ministro Celso de Mello, decano do Supremo Tribunal Federal (STF) a favor dos embargos infringentes. Além das postagens no primeiro dia de coleta, que antecede o dia da votação, o grande volume de links postados foi concentrado nos dois dias seguintes, com links para páginas que transmitiam a votação ao vivo e outros portais de notícias e também com postagens noticiando o resultado da votação e seus respectivos impactos.

Os datasets aqui identificados são apenas os que serão utilizados na análise. Conforme o explicado anteriormente, apenas os datasets filtrados e seus conjuntos de interesse serão utilizados, havendo o descarte dos datasets preliminares. Os grafos descritos nesta seção foram gerados a partir dos datasets filtrados, e os conjuntos de interesse de cada grafo são equivalentes aos conjuntos de interesse obtidos a partir dos datasets filtrados de cada um dos três termos.

Nos grafos apresentados a seguir, os vértices representam os usuários do *Twitter* e as arestas representam as postagens, ou *tweets*. As arestas de cor cinza são as postagens que não contém URLs, e as arestas coloridas representam as postagens que contém URLs. É por este motivo que os grafos dos conjuntos de interesse não possuem nenhuma aresta da cor cinza. Cada cor diferente de aresta representa um URL diferente.

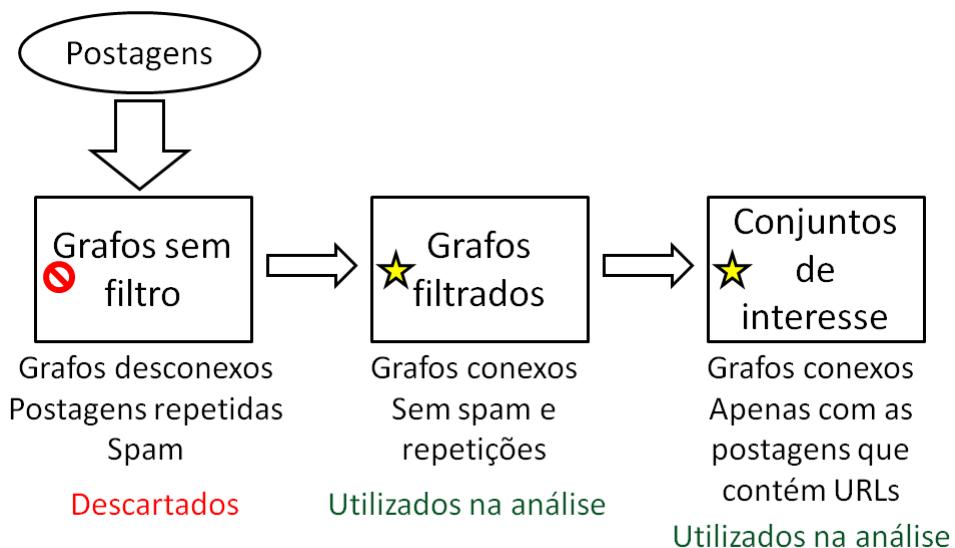


Figura 3.6: Processo de obtenção das amostras

### 3.3.1 Grafo STF

O **grafo STF** corresponde ao dataset filtrado da coleta resultante do termo 'STF' e possui 5719 vértices e 14308 arestas, sendo 10206 únicas e 4102 duplicatas. O diâmetro do grafo é 13, a densidade é de aproximadamente 0,00032 e a distância média entre os vértices é de aproximadamente 4,65.

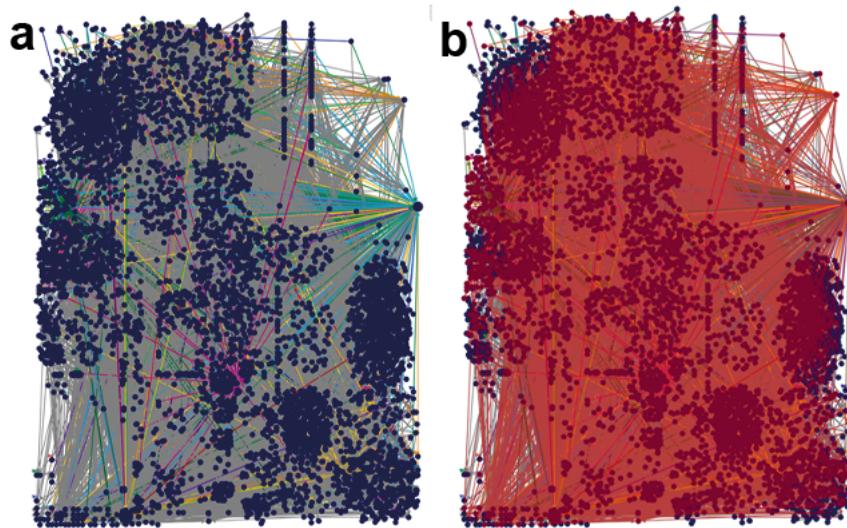


Figura 3.7: a) Grafo STF. b) Conjunto de interesse STF destacado em vermelho.

O conjunto de interesse STF, que é um subgrafo do grafo STF, possui 1384 vértices e um total de 2016 arestas, dentre as quais 1545 são únicas e 471 possuem duplicatas. O diâmetro do grafo do conjunto de interesse é 16, valor superior ao do grafo que o originou. A distância média entre os vértices do grafo é de aproximadamente 4,66 e a densidade é de aproximadamente 0,001.

### 3.3.2 Grafo AP470

O grafo AP470 corresponde ao dataset filtrado da coleta resultante do termo 'AP470' e possui um total de 437 vértices e 1063 arestas, sendo 761 únicas e 302 com duplicatas. O diâmetro do grafo é 7, a distância média entre os vértices é de aproximadamente 3,2 e a densidade é de aproximadamente 0,004.

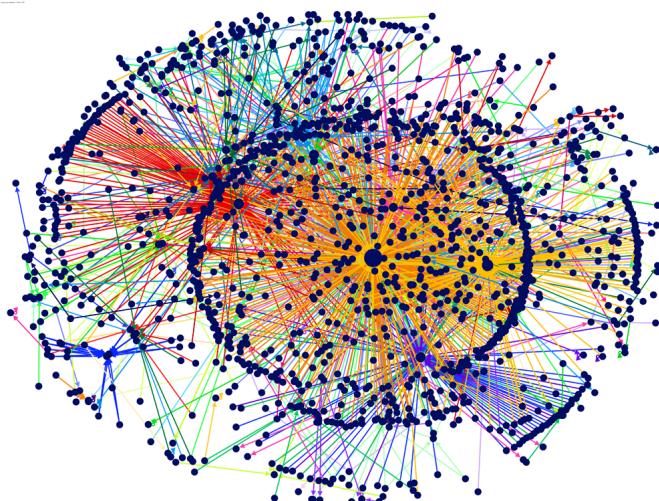


Figura 3.8: Conjunto de interesse STF

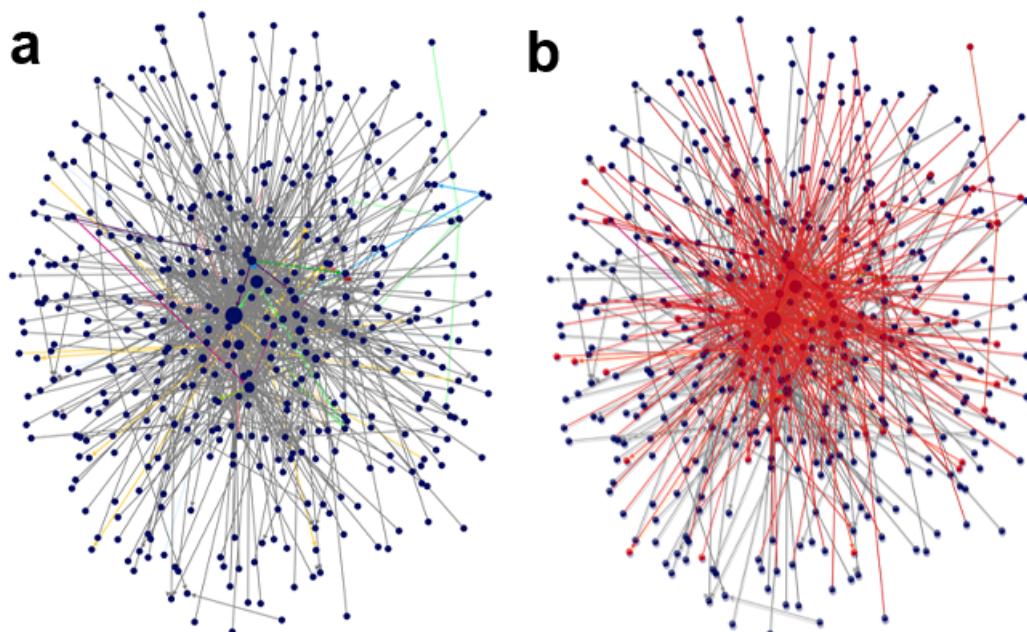


Figura 3.9: a) Grafo AP470. b) Conjunto de interesse AP470 destacado em vermelho.

O conjunto de interesse AP470, que vem a ser um subgrafo do grafo AP470, contém 58 vértices e 92 arestas, sendo 81 únicas e 11 com duplicatas. O diâmetro deste grafo é 2, a distância média entre os vértices é de aproximadamente 1,8 e a densidade do grafo é de aproximadamente 0,48.

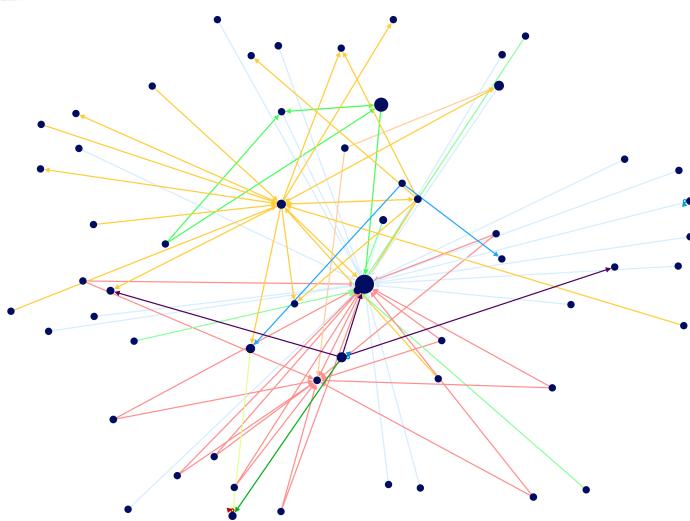


Figura 3.10: Conjunto de interesse AP470

### 3.3.3 Grafo #CelsoDeMello

O grafo #CelsoDeMello é equivalente ao dataset filtrado da coleta resultante do termo '#CelsoDeMello' e contém 543 vértices e 1182 arestas. O diâmetro do grafo é 10, a densidade do grafo é de aproximadamente 0,003 e a distância média entre os vértices é aproximadamente 5.

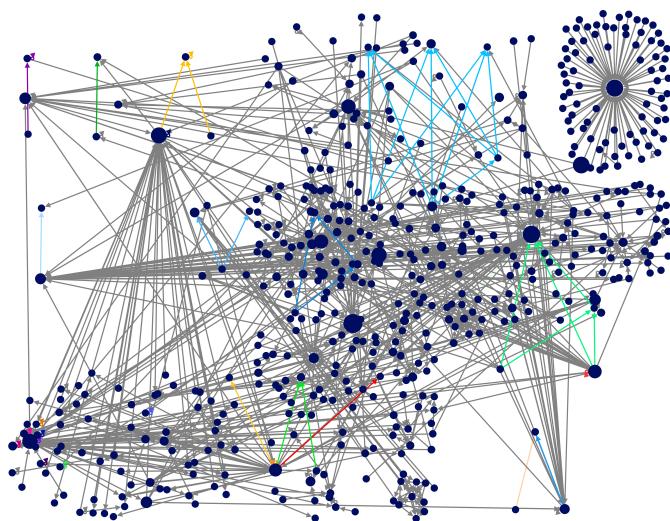


Figura 3.11: Grafo #CelsoDeMello

Utilizando o mesmo critério para a obtenção do conjunto de interesse utilizado nos outros grafos, o conjunto de interesse #CelsoDeMello contém 6 vértices e 17 arestas, sendo 3 únicas e 14 com duplicatas, e todas as arestas do conjunto contém a mesma URL. Este conjunto é

substancialmente menor que os outros anteriormente apresentados.

É possível observar no grafo da figura 3.12 que suas componentes conexas são todas pequenas, e isso faz com que as métricas dos grafos e vértices obtidas a partir deste conjunto de interesse destoem das obtidas a partir dos outros grafos. Por esses motivos, é preferível não incluir o grafo #CelsoDeMello nem seu conjunto de interesse nas análises. Apesar de o conjunto de interesse #CelsoDeMello ser o único que destoa dos demais, manter o grafo #CelsoDeMello nas análises e não ter nenhum conjunto de interesse com o qual compará-lo ocasionaria em uma inconsistência no processo de análise, que, a princípio, consiste na comparação entre os grafos e seus respectivos conjuntos de interesse.

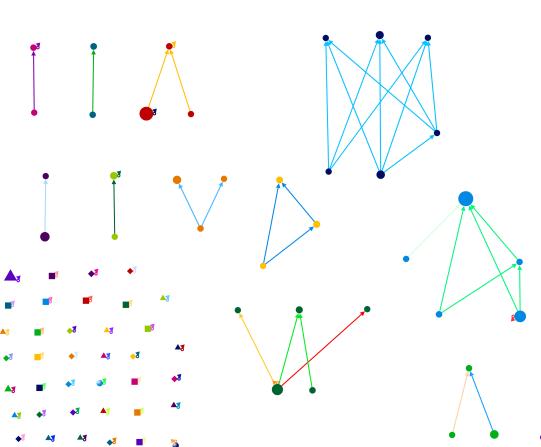


Figura 3.12: Grafo desconexo, contendo apenas as interações do grafo #CelsoDeMello que continham endereços da web nas postagens

Ao longo deste capítulo, foram apresentados os instrumentos utilizados para a coleta, organização e análise dos dados, além do resultado preliminar da coleta, que contém os grafos a serem analisados no próximo capítulo. Todavia, antes disso, algumas considerações são necessárias.

Apesar de a realização de todas as coletas ter ocorrido no mesmo período de tempo e na mesma quantidade de vezes, percebe-se a diferença no tamanho dos três grafos obtidos e como isso impactou os conjuntos de interesse em cada grafo. Isso se deve, principalmente, aos termos escolhidos para guiar a pesquisa. Ao escolher o termo 'STF', foram coletadas todas as ocorrências, inclusive *hashtags* e menções ao perfil do @STF\_oficial. Em muitas das vezes que mencionavam o perfil do STF, o nome de usuário do perfil era a única ocorrência do termo 'STF' na postagem.

A quantidade de ocorrências do termo 'AP470' foi menor que a do termo 'STF', mesmo realizando a coleta sem a restrição da *hashtag*. Durante os dias da coleta e até mesmo alguns

dias depois, uma busca feita no *Twitter* pelo termo 'STF' retornava resultados novos em questão de segundos, devido ao grande fluxo de postagens contendo o termo. Ao buscar pelo termo 'AP470', observava-se um intervalo de tempo maior entre as postagens. Ambos os termos tiveram um grande volume de mensagens durante o período de coleta, mas o termo 'STF' estava muito mais em evidência do que o termo 'AP470', apesar de a votação do Ministro envolver diretamente a Ação Penal 470.

Além dessas diferenças na quantidade de ocorrências de cada termo, o que causou um impacto direto na obtenção de cada um dos grafos, muitas outras foram observadas durante a análise dos grafos. O processo de análise e os resultados obtidos serão descritos a seguir.

## 4. Análise

Este capítulo discorre sobre a análise dos quatro grafos apresentados no capítulo anterior. Deu-se início à análise a partir da identificação dos vértices-chave de cada grafo utilizando as métricas de vértice. Posteriormente, foram realizadas comparações entre os grafos e seus respectivos conjuntos de interesse e também comparações entre os grafos. Inicialmente, será detalhado o processo de identificação dos vértices-chave. Nos grafos apresentados anteriormente, cada aresta representa uma postagem diferente e cada vértice representa um perfil de usuário, ou simplesmente um usuário do *Twitter*.

### 4.1 Identificação dos vértices-chave

A análise inicial, com o objetivo de identificar os vértices-chave em cada um dos grafos, foi feita a partir das métricas de vértice calculadas pelo *NodeXL* para os grafos STF e AP470 e seus respectivos conjuntos de interesse. Para cada um desses quatro grafos, duas tabelas foram elaboradas. As tabelas se encontram no **Apêndice**, ao final do trabalho. As sete colunas representam, cada uma, uma métrica diferente, em ordem: grau de entrada, grau de saída, centralidade de intermediação, centralidade de proximidade, centralidade de autovetor, *PageRank* e coeficiente de agrupamento.

A primeira tabela contém duas linhas, sendo que cada uma deve mostrar os maiores e menores valores que cada métrica assume naquele grafo. Os valores representados na mesma linha não necessariamente correspondem aos mesmos vértices. A primeira linha contém os maiores valores encontrados para cada métrica dos vértices no grafo, enquanto a segunda linha contém os menores.

A segunda tabela é montada a partir da primeira. Uma das principais diferenças entre a primeira tabela e a segunda é que esta visa identificar os vértices contendo os maiores ou menores valores de cada métrica, também levando em consideração os valores obtidos para as outras métricas daquele mesmo vértice. Portanto, cada linha da segunda tabela contém os valores de cada métrica de um mesmo vértice do grafo. Deste modo, é possível observar a

correlação entre as métricas e identificar potenciais vértices-chave para o estudo.

Outro ponto importante da segunda tabela é que ela foi elaborada de forma a evitar repetições de um mesmo vértice em diversas linhas. Por exemplo, se um vértice possui o maior grau de entrada e também o maior *PageRank*, tais valores serão agrupados em uma mesma linha, para evitar redundâncias.

Tal critério revelou-se útil após a constatação de que em todas as tabelas montadas, vértices com os maiores graus de entrada também possuíam os maiores valores em centralidade de intermediação, centralidade de proximidade, centralidade de autovetor e *PageRank*. Por esses motivos, são destacadas apenas as linhas cujos vértices possuem maior grau de entrada, maior grau de saída e maior coeficiente de agrupamento, que são as três características que não se sobrepõem ao se levar em consideração os maiores valores que tais métricas alcançam.

O mesmo critério foi utilizado ao levar em consideração os menores valores que cada uma dessas métricas poderia assumir, mas neste caso as métricas em que um mesmo vértice não consegue obter o menor valor do grafo simultaneamente são grau de entrada, grau de saída e *PageRank*. Apesar da aparente complexidade, tal organização simplificou a tabela e facilitou a visualização das correlações entre os valores. As tabelas encontram-se no apêndice do trabalho.

Foram considerados vértices-chave de cada grafo os que apresentaram os maiores valores nas métricas de vértice, exceto pelo maior coeficiente de agrupamento. Os vértices que apresentaram os menores valores de métricas foram destacados na tabela para a realização de possíveis comparações entre os grafos. O principal motivo para a não identificação desses vértices como vértices-chave é que os menores valores assumidos por uma métrica tendem a aparecer em diversos vértices de um mesmo grafo, o que torna delicada a tarefa de identificar um vértice que represente adequadamente todos os outros do grupo. Os vértices com maiores coeficientes de agrupamento foram desconsiderados por possuírem baixa conectividade. Apesar de coeficiente de agrupamento mais elevado, tais vértices não possuem muitas conexões, o que torna a ocorrência de tríades mais provável. A seguir, serão identificados os vértices-chave de cada um dos grafos e de seus respectivos conjuntos de interesse.

De acordo com os critérios estabelecidos anteriormente, os vértices-chave do grafo STF são @STF\_oficial, perfil oficial do STF no *Twitter*, que apresenta maior grau de entrada, centralidade de intermediação, centralidade de autovetor e *PageRank* e @faxinanopoder, com o maior grau de saída. A centralidade de proximidade não foi levada em consideração no grafo STF por ter apresentado valores igualmente inexpressivos em todos os vértices do grafo.

No conjunto de interesse STF, o vértice @estadao, que é o perfil do Estadão, a versão online

do jornal O Estado de São Paulo, possui o maior grau de entrada e também as maiores centralidades de intermediação e autovetor, além do maior *PageRank*. O vértice @eljabberwocky apresenta o maior grau de saída e a maior centralidade de proximidade.

No grafo AP470, o vértice @stanleyburburin assume os maiores valores de grau de entrada, centralidade de intermediação, centralidade de autovetor e *PageRank* e @jprcampos é o vértice com o maior grau de saída.

No conjunto de interesse AP470, @stanleyburburin é novamente o vértice com os maiores valores de grau de entrada, centralidade de intermediação, centralidade de autovetor e *PageRank*, além de ser o vértice que possui o maior valor de centralidade de proximidade, que, no grafo AP470 era aproximadamente o mesmo para todos os vértices. O vértice @ptnosenado se destaca por apresentar o maior grau de saída.

Com os vértices-chave definidos, a próxima etapa é localizá-los nos outros três grafos e observar suas características e métricas. O objetivo desta etapa é averiguar se a presença dos vértices-chave de um grafo específico será mantida nos outros grafos.

## 4.2 Participação dos vértices-chave nos outros grafos

Para avaliar a participação dos vértices-chave na conversa como um todo, os vértices anteriormente identificados serão localizados em cada um dos outros três grafos.

Observa-se que o vértice @stanleyburburin é o único vértice-chave que se destacou em mais de um grafo, contendo as maiores medidas de centralidade e também sendo considerado o vértice mais relevante tanto no grafo AP470 quanto em seu conjunto de interesse, por possuir *PageRank* mais elevado que os demais. O mesmo também aparece no grafo STF e também em seu conjunto de interesse, sempre entre os vértices com os maiores graus de entrada. Além disso, @stanleyburburin é um dos poucos vértices que possuem valores expressivos tanto do grau de entrada quanto de saída, o que indica o engajamento do vértice em ambos os sentidos das interações de todos os grafos analisados.

Já o vértice @STF\_oficial, que possui o maior grau de entrada no grafo STF, além de não ser vértice-chave no conjunto de interesse STF, também não assume nenhuma posição de destaque no grafo AP470 ou no conjunto de interesse AP470. Além disso, o mesmo comportamento observado pelo perfil oficial do STF no grafo STF é observado em todos os outros grafos, já que o grau de saída do vértice @STF\_oficial sempre assume valores baixos. Como o perfil do STF não menciona nenhum outro usuário, os inexpressivos valores de grau de saída que ele

assume são oriundos de laços no grafo. O mesmo ocorre com o vértice @estadao, que, apesar de disseminar grande quantidade de notícias tanto no grafo STF quanto em seu conjunto de interesse, mal participa do grafo AP470 e não está presente no conjunto de interesse AP470. Os vértices @faxinanopoder e @eljabberwocky não foram encontrados no grafo AP470, o que implica na não participação dos mesmos no conjunto de interesse AP470.

O vértice @jprcampos, um dos vértices-chave do grafo AP470, também obteve certo destaque no grafo STF, com um dos maiores graus de saída observados. Apesar de estar entre os vértices mais importantes de ambos os grafos STF e AP470, ele não foi encontrado no conjunto de interesse AP470 e sua participação foi mínima no conjunto de interesse STF. O vértice @ptnosenado obteve pouco destaque tanto no grafo STF quanto em seu conjunto de interesse, mas no grafo AP470, obteve métricas semelhantes às do vértice @jprcampos no mesmo grafo.

A partir das observações feitas acima sobre os vértices-chave nos quatro grafos, é possível obter uma ideia de como os usuários brasileiros utilizam o *Twitter* para acompanhar e discutir acontecimentos do cenário político, além de como perfis de maior alcance nacional, como o perfil oficial do STF e o do Estadão, interagem com o público do *Twitter*. Tais considerações serão feitas na próxima etapa da análise, em que todos os grafos serão comparados entre si, utilizando métricas de grafo e também as considerações feitas a partir da observação da participação dos vértices-chave.

## 4.3 Análise comparativa dos grafos

A análise comparativa dos grafos consiste essencialmente na comparação entre os valores das métricas de grafo e também pelo que foi observado na participação dos vértices-chave em todos os quatro grafos. Inicialmente, serão feitas algumas considerações sobre o grafo STF e seu respectivo conjunto de interesse, seguidas pelas considerações referentes ao grafo e conjunto de interesse AP470. Assim como as métricas de vértice, as métricas de grafo aqui apresentadas foram calculadas pelo *NodeXL*.

### 4.3.1 Grafo e Conjunto de Interesse STF

Ao analisar as tabelas referentes ao grafo STF e seu conjunto de interesse, percebe-se que a métrica *PageRank* assume seu maior valor no conjunto de interesse e não no grafo. Como o grafo STF é muito maior que seu conjunto de interesse, uma possível explicação para o ocorrido é que, com a diminuição do tamanho do grafo, a importância de alguns vértices aumentou,

ocasionando o aumento nos valores dessas métricas.

Também é importante ressaltar que o vértice com o *PageRank* mais alto do grafo STF não é o mesmo do conjunto de interesse. Enquanto este é responsável por disseminar links de notícias, o outro é apenas mencionado várias vezes, mas não porque outros usuários estão repassando suas postagens pela rede.

O perfil do Estadão possui mais *retweets*, o que significa que seus vizinhos no grafo tinham o objetivo de ler e disseminar notícias sobre a votação. Ao observar o grafo do conjunto de interesse do STF, é perceptível que as arestas apontando para o vértice que representa o @estadao, que é o maior vértice, no centro do grafo, possuem diversas cores. Arestas das mesmas cores contém as mesmas URLs nas postagens, e arestas de cores diferentes representam URLs diferentes. O perfil do Estadão foi o maior referencial de notícias para o conjunto de interesse, tendo muitos links para diferentes notícias disseminados pela rede.

Em contrapartida, o perfil oficial do STF foi alvo de diversas postagens que tinham o objetivo de criticar a decisão do Ministro Celso de Mello e também de alguns *retweets* sobre o momento em que a decisão foi tomada. Na maioria das postagens do grafo STF, principalmente as direcionadas ao perfil do @STF\_oficial, é perceptível o descontentamento dos usuários com a decisão tomada pelo Ministro. A maioria desses comentários, principalmente os que mencionavam o perfil oficial do STF, são feitos por usuários que não se engajaram em discussões no grafo. Ao analisar o grau de entrada e de saída desses vértices que mencionaram o perfil do STF com comentários como os listados acima, observa-se que a maioria desses valores são baixos. Muitos desses vértices só aparecem no grafo por terem mencionado @STF\_oficial, sendo esta a única conexão de tais vértices com outros no grafo. Apesar desses vértices mencionarem o vértice mais central do grafo, o comportamento deles não difere muito dos vértices que não interagiam com nenhum outro no dataset preliminar.

Todavia, o perfil do STF, mesmo sendo o vértice mais central do grafo, também não engaja em nenhuma conversa. Apesar de ter o grau de entrada mais elevado entre milhares de vértices, seu grau de saída é muito baixo, o que significa que a reciprocidade da conta com a sua audiência é praticamente inexistente. Esta constatação ajuda a entender o comportamento de outras autoridades políticas brasileiras presentes no *Twitter*.

Apesar de o *Twitter* ser um canal que possibilita a interação bilateral sem a necessidade de um vínculo recíproco, ou seja, possibilita que dois usuários que não seguem um ao outro ainda assim troquem mensagens em algum momento, ele também passa uma falsa impressão de interação e proximidade quando, na realidade, apenas reforça a distância entre aquelas conexões. Neste caso, os usuários utilizam o que acredita-se ser um canal direto de comunicação com o

Supremo Tribunal Federal para expressar suas opiniões e ficam sem resposta. Vale ressaltar que não é esperado que um usuário com um grande volume de seguidores e mensagens responda a todas elas, mas no caso do perfil do STF a falta de interesse de interagir com uma quantidade mínima de seguidores é notória.

Ao comparar o perfil oficial do STF com o perfil @faxinanopoder, vértice que obteve o maior grau de saída no grafo STF, percebe-se como este usou o *Twitter* de uma forma diferente. Apesar de possuir um grau de entrada 103, que é significativamente menor que o do perfil do STF, seu grau de saída corresponde a quase 40% dessa quantidade de menções e *retweets*. Apesar de isso não significar que o perfil necessariamente interagiu com os usuários que o mencionaram, o que não foi o caso, mostra que tal usuário estava presente em ambos os lados das interações. Todavia, ao observar o comportamento dos vértices do grafo como um todo, constata-se que são poucos os usuários com um papel semelhante ao do @faxinanopoder na rede.

Ao comparar os números das tabelas do grafo e do conjunto de interesse, é possível perceber que métricas que enfatizam a participação de cada vértice em conexões no grafo como um todo, como a centralidade de proximidade e a centralidade de autovetor, manifestam seus maiores valores no conjunto de interesse, que é um subgrafo do grafo STF.

### 4.3.2 Grafo e Conjunto de Interesse AP470

Ao analisar as tabelas referentes ao grafo AP470 e a seu conjunto de interesse, são observadas algumas variações nas métricas dos vértices destacados. Tais variações são semelhantes às observadas no grafo STF e em seu conjunto de interesse. Por exemplo, o que ocorre com os valores da centralidade de proximidade e a centralidade de autovetor.

Uma possível justificativa para tais resultados é o fato de o conjunto de interesse conter uma quantidade consideravelmente menor de vértices do que o do grafo. No grafo, inclusive, o maior valor encontrado para a centralidade de proximidade é inferior ao menor valor encontrado no conjunto de interesse. Isso mostra que até o vértice menos central do conjunto de interesse poderia interagir mais rapidamente com qualquer outro vértice do que o vértice mais central do grafo.

A diferença observada entre outras métricas como grau de entrada, centralidade de intermediação e *PageRank* é justificada de forma mais direta. Ao contrário do ocorrido com o grafo STF e seu respectivo conjunto de interesse, a métrica *PageRank* apresentou seu maior valor no grafo e não no conjunto de interesse. Tais valores possuem uma forte tendência a acompanhar

o tamanho do grafo, e é exatamente o que acontece nas amostras observadas.

Outra ocorrência que se destaca na tabela 2 do conjunto de interesse é que o vértice com o menor *PageRank* obtém maiores valores de centralidade do que vértices com maior *PageRank*. Isso é justificado ao verificar as conexões de cada um desses vértices no grafo. Os vértices com menor centralidade de proximidade estão ligados a vértices menos centrais, o que significa que tais vértices, mesmo sendo melhor ranqueados que outros, não necessariamente alcançam outros vértices com menos passos.

O vértice com o maior grau de entrada do grafo AP470 é o mesmo do conjunto de interesse. Percebe-se que as métricas de centralidade de proximidade e centralidade de autovetor apresentam valores maiores no conjunto de interesse do que no grafo.

No conjunto de interesse, há uma convergência maior entre os vértices e isso é perceptível através dessas duas métricas. A densidade do conjunto de interesse AP470 é de aproximadamente 0,48, enquanto a do grafo AP470 é aproximadamente 0,004. Outra métrica do grafo que mostra isso é a distância média entre os vértices, que no conjunto de interesse é aproximadamente 1,8 e aproximadamente 3,2 no grafo.

Ao relacionar as métricas com os respectivos usuários e suas interações no grafo como um todo, o grafo AP470 e seu conjunto de interesse se revelam muito diferentes do que foi observado no grafo STF e em seu conjunto de interesse. Os usuários centrais, apesar de não possuírem graus de entrada tão altos quanto o perfil oficial do STF, interagem mais com os outros usuários e são encontrados com facilidade em ambos os lados das interações. Apesar de ainda serem poucos os vértices com essas características, seus altos valores de centralidade mostram que a audiência ao redor os valoriza mais do que perfis de jornais online e de autoridades políticas. No grafo AP470 e em seu conjunto de interesse, o foco principal dos usuários que o integram é discutir o assunto com menos superficialidade.

As mensagens nesses conjuntos possuem um caráter menos impulsivo que o que foi observado no grafo STF, e, portanto, são mais longas. A intenção desses usuários não era apenas de expor suas opiniões, mas sim de interagir com os outros usuários que postavam sobre o mesmo assunto. Mais do que termos diferentes, os grafos STF e AP470 e seus respectivos conjuntos de interesse representam grupos e perspectivas diferentes. Prova disso são as menções feitas ao perfil oficial do STF pelos usuários presentes no grafo AP470. Enquanto no grafo STF o @STF\_oficial é mencionado 844 vezes, no grafo AP470 são feitas apenas onze menções ao perfil oficial.

Algo importante observado nesta análise é que os perfis com valores muito expressivos de

grau de entrada tendem a possuir baixos valores de graus de saída e, apesar de possuírem altos graus de centralidade, se situam apenas em um dos lados das interações. Tal tendência é exemplificada pelos perfis do STF e do Estadão. O grau de saída, todavia, não acompanha a mesma tendência. Usuários cujos valores de grau de saída são mais expressivos não acompanham nenhum padrão específico de relação com o grau de entrada, pois geralmente as menções que tais usuários receberão dependem de diversos fatores. Entre alguns destes fatores, estão a qualidade das postagens e a predisposição do usuário que foi mencionado primeiro de responder àquela postagem, que também depende do quanto um usuário quer engajar em uma discussão.

Além disso, constata-se que o comportamento da maioria dos vértices do grafo influencia na definição dos vértices-chave. No grafo STF, por exemplo, cuja maior parte do conteúdo é de postagens de conteúdo vazio, mensagens curtas, de teor mais intenso e muitas vezes até ofensivas, quase todas direcionadas ao Supremo Tribunal Federal ou aos seus Ministros, não é surpresa que o perfil do STF seja o vértice com mais mensagens recebidas, apesar de não responder nenhuma. Já no grafo AP470, os usuários interagiam mais entre si e engajavam em discussões mais substanciais, o que justifica o destaque do vértice @stanleyburburin, que esteve constantemente envolvido nas discussões em todos os grafos.

## 5. Conclusão

A partir de três amostras, que juntas totalizaram mais de 16.000 postagens do *Twitter*, cerca de 15.000 foram analisadas ao longo deste trabalho. Ao longo deste capítulo, serão feitas as considerações finais sobre este trabalho, como seus resultados, limitações e trabalhos futuros.

Ao longo da análise das redes, foi observado que algumas métricas de vértice tendem a ser diretamente proporcionais às outras, ao levar em consideração vértices de um mesmo grafo. Por exemplo, em todos os quatro grafos analisados, os vértices que possuem o maior grau de entrada também são os que possuem maior centralidade de intermediação, centralidade de autovetor e PageRank, isto é, o vértice que possui o maior valor em qualquer uma dessas métricas no grafo tende a possuir os maiores valores nas outras métricas supracitadas. Porém, essa proporcionalidade só é mantida no grafo e só é válida para os maiores valores das métricas. Vértices cujas métricas assumem valores intermediários tendem a não obedecer proporções ou padrões específicos, o que também acontece ao comparar as métricas de vértices pertencentes a grafos distintos.

Também foi observado que o coeficiente de agrupamento possui um comportamento inversamente proporcional ao grau do vértice, isto é, quanto maior o grau de um vértice, menor tende a ser seu coeficiente de agrupamento. Principalmente nos quatro grafos analisados, em que os vértices são bastante dispersos entre si, e a proporção de tríades formadas é baixa. A centralidade de proximidade assume seus maiores valores absolutos nos conjuntos de interesse, que são subgrafos das duas amostras maiores.

Sobre o comportamento dos usuários envolvidos nas discussões, foi observado que o vértice com o maior grau de entrada tende a ditar o comportamento dos outros vértices do grafo, de forma direta ou indireta. Por exemplo, no grafo STF, cujo vértice com o maior grau de entrada era o perfil oficial do STF, observou-se que o teor das postagens no grafo em geral refletia o teor impulsivo das mensagens direcionadas ao STF. Alguns pequenos grupos dentro do grafo se comportavam de maneira diferente, mas a grande tendência no grafo STF não era discutir os impactos do voto do Ministro ou disseminar notícias, apenas transmitir as impressões de cada usuário sobre o ocorrido, sejam favoráveis ou desfavoráveis.

Já no conjunto de interesse STF, em que o vértice com o maior grau de entrada era o perfil do Estadão, a tendência apresentada era de postar links com notícias que anunciam não apenas o resultado do voto, mas que também explicavam os impactos do resultado no julgamento e as possíveis reduções de pena. Algumas postagens também continham pequenos comentários, alguns ainda refletindo o caráter impulsivo oriundo do grafo STF, que originou esse conjunto de interesse.

Já no grafo AP470 e em seu conjunto de interesse, o teor das mensagens é menos impulsivo, as mensagens são mais longas e sugerem a possibilidade de discussão. É possível perceber através da análise que o comportamento dos atores nestes grafos reflete o comportamento do vértice mais popular, o usuário @stanleyburburin, um dos poucos vértices presentes em todos os quatro grafos analisados e sempre engajado em ambos os lados da conversa. Este vértice não apenas era frequentemente mencionado, mas também mencionava outros usuários e engajava em discussões frequentemente. Tal tendência é um indicativo de que o vértice-chave identificado pelas métricas de vértice pode ser um ponto de partida útil para a análise do comportamento dos vértices de um grafo, já que o que acontece ao redor daquele vértice possui uma tendência a se propagar para os outros vértices do grafo.

Também é importante ressaltar que este estudo contém certas limitações, já que foi executado em um curto intervalo de tempo e contém uma quantidade expressiva de dados. Dos mais de 17.000 postagens coletadas, pouco mais de 16.000 foram submetidas à análise durante um período de dois meses e meio. Além disso, como o idioma predominante das postagens foi o português brasileiro, não foi possível fazer análises automatizadas procurando por opiniões positivas ou negativas com o auxílio do NodeXL, que só é capaz de fazer isso com postagens em inglês. Além disso, pela grande quantidade de dados analisados e pela própria estrutura de organização dos mesmos, não era possível identificar todos os vértices que estavam presentes em mais de um grafo. Devido a essas limitações e à própria restrição de tempo, a análise manual foi utilizada com frequência.

Apesar de a quantidade de dados coletados ser expressiva, tanto para o tempo de coleta quanto de análise, ela poderia ser ainda maior caso mais termos fossem coletados. Neste caso, as tendências observadas poderiam ser confirmadas ou até mesmo outras tendências pouco perceptíveis nestes quatro grafos poderiam surgir. Todavia, devido à complexidade de se realizar coletas simultâneas para mais de três termos em intervalos de tempo regulares, a coleta foi restrita à apenas três termos, sendo eles 'STF', 'AP470' e '#CelsoDeMello'. Caso mais amostras fossem coletadas, as possibilidades de comparação também aumentariam, assim como de critérios para a obtenção de subgrafos.

Sugestões para trabalhos futuros envolvendo este estudo incluem a realização de análise de sentimento das postagens presentes nas amostras, utilização de bases de dados maiores, seja pela extensão do período de coleta ou pelo aumento na quantidade de termos buscados e a expansão para outros acontecimentos no cenário político brasileiro. Tais acontecimentos podem ser inicialmente, de impacto nacional, e posteriormente, separados por regiões do país ou até mesmo por estados, para fins de comparação.

## 6. Referências

- BACKSTROM, L.; KLEINBERG, J. **Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook.** CSCW?14, 2013.
- BERTI, O. M. C. **O Twitter e suas interfaces com o regional. Um estudo da produção noticiosa regional e sua retroalimentação informacional no Piauí.** CONFIBERCOM, São Paulo, 2011
- BRANDES, U.; PICH, C. **Centrality Estimation in Large Networks.** Department of computer and Information Science, university of Kontanz. August 18, 2006.
- BURGE, K.; COMM, J. **O poder do Twitter.** São Paulo. Ed. Gente, 2009.
- CASSAES, D.; GARCIA, R. T. **Produção e consumo de notícia: o Twitter enquanto ferramenta jornalística.** Revista Lumen Et Virtus, vol II, nº 4, maio de 2011.
- CHA, M.; HADDADI, H.; BENEVENUTO, F.; GUMMADI, K. P. **Measuring User Influence in Twitter: The Million Follower Fallacy.** Association for the Advancement of Artificial Intelligence, 2010.
- CHO, J.; ROY, S. **Impact of search engines on page popularity.** 13th international conference on World Wide Web (pp. 20-29). ACM, 2004.
- HANSEN, D. L.; SHNEIDERMAN, B.; SMITH, M. A. **Analyzing Social Media Networks with NodeXL: Insights from a Connected World.** Morgan Kaufmann, 2011.
- HUBERMAN, B. A.; ROMERO, D. M.; WU, F. **Social Networks that Matter: Twitter Under the Microscope.** Disponível em SSRN 1313405, 2008.
- JANSEN, B. J.; ZHANG, M.; SOBEL, K.; CHOWDURY, A. **Twitter power: Tweets as electronic word of mouth.** Journal of the American society for information science and technology, 60(11), 2169-2188, 2009.
- JARAMILLO, A. M. **Twitter para todos: su negocio em 140 caracteres.** Bogotá: Vergara, 2010.

- KRISHNAMURTHY, B.; GILL, P.; ARLITT, M. **A Few Chirps About Twitter.** WOSN'08, 2008.
- NEWMAN, M. E. J. **Networks: An Introduction.** Oxford University Press, 2010.
- SERRAT, O. **Social Network Analysis.** Washington, DC: Asian Development Bank, 2010. Disponível em: <http://digitalcommons.ilr.cornell.edu/intl/206>
- SUZUKI, H. T.; RIBEIRO, C. H. C. **Estudo Empírico do Fenômeno Small World em Redes Sociais.** Anais do 13º Encontro de Iniciação Científica e Pós-graduação do ITA - XIII ENCITA, São José dos Campos, 2007.
- TUMASJAN, A., SPRENGER, T. O., SANDNER, P. G., WELPE, I. M. **Predicting Elections With Twitter: What 140 Characters Reveal about Political Sentiment.** Association for the Advancement of Artificial Intelligence, 2010.
- WANG, Z.; SCAGLIONE, A.; THOMAS, R. J. **Electrical centrality measures for electric power grid vulnerability analysis.** Decision and Control (CDC), 49th IEEE Conference, 5792-5797, 2010.
- WASSERMAN, S.; FAUST, K. **Social Networks Analysis: Methods and Applications.** Cambridge: Cambridge University Press. 1994.
- WILLIAMS, C. B., GULATI, G. J. **What is a Social Network Worth? Facebook and Vote Share in the 2008 Presidential Primaries.** Annual Meeting of the American Political Science Association. 2008

## 7. Apêndice

### Apêndice A - tabelas do grafo STF e conjunto de interesse STF

#### Grafo STF

	Grau de entrada (g.e.)	Grau de saída (g.s.)	Centralidade de Intermediação	Centralidade de Proximidade	Centralidade de Autovetor	PageRank	Coeficiente de Agrupamento
Maior	844	39	13833493,230	0	0,024	188,210	1
Menor	0	0	0	0	0	0,275	0

Tabela 1 do grafo STF

	Grau de entrada	Grau de saída	Centralidade de Intermediação	Centralidade de Proximidade	Centralidade de Autovetor	PageRank	Coeficiente de Agrupamento
Maior g.e.	844	1	13833493,230	0	0,024	188,210	0,001
Maior g.s.	103	39	1446640,731	0	0,004	24,937	0,015
Maior coef. agr.	3	0	0	0	0	0,717	1
Menor g.e.	0	1	0	0	0	0,310	0
Menor g.s.	1	0	0	0	0	0,310	0
Menor PageRank	0	1	0	0	0	0,275	0

Tabela 2 do grafo STF

## Conjunto de interesse STF

	Grau de entrada	Grau de saída	Centralidade de Intermediação	Centralidade de Proximidade	Centralidade de Autovetor	PageRank	Coeficiente de Agrupamento
Maior	461	8	1258805,442	0,143	0,044	194,623	1
Menor	0	0	0	0	0	0,415	0

Tabela 1 do conjunto de interesse STF

	Grau de entrada	Grau de saída	Centralidade de Intermediação	Centralidade de Proximidade	Centralidade de Autovetor	PageRank	Coeficiente de Agrupamento
Maior g.e.	461	1	1258805,442	0	0,044	194,623	0
Maior g.s.	2	8	42	0,143	0	3,986	0
Maior coef. agr.	0	2	0	0	0	0,567	1
Menor g.e.	0	1	0	0,026	0	0,557	0
Menor g.s.	1	0	0	0	0	0,522	0
Menor PageRank	1	0	0	0	0	0,415	0

Tabela 2 do conjunto de interesse STF

## Apêndice B - tabelas do grafo AP470 e conjunto de interesse AP470

### Grafo AP470

	Grau de entrada	Grau de saída	Centralidade de Intermediação	Centralidade de Proximidade	Centralidade de Autovetor	PageRank	Coeficiente de Agrupamento
Maior	153	16	104975,041	0,001	0,049	36,166	1
Menor	0	0	0	0,000	0,001	0,330	0

Tabela 1 do grafo AP470

	Grau de entrada	Grau de saída	Centralidade de Intermediação	Centralidade de Proximidade	Centralidade de Autovetor	PageRank	Coeficiente de Agrupamento
Maior g.e.	153	14	104975,041	0,001	0,049	36,166	0,007
Maior g.s.	8	16	8750,433	0,001	0,009	4,973	0,053
Maior coef. agr.	0	2	0	0,001	0,001	0,564	1
Menor g.e.	0	1	0	0,001	0,001	0,404	0
Menor g.s.	1	0	0	0,001	0,001	0,334	0
Menor PageRank	1	0	0	0,001	0	0,330	0

Tabela 2 do grafo AP470

## Conjunto de interesse AP470

	Grau de entrada	Grau de saída	Centralidade de Intermediação	Centralidade de Proximidade	Centralidade de Autovetor	PageRank	Coeficiente de Agrupamento
Maior	32	11	2495,667	0,011	0,124	10,767	1
Menor	0	0	0	0,004	0	0,427	0

Tabela 1 do conjunto de interesse AP470

	Grau de entrada	Grau de saída	Centralidade de Intermediação	Centralidade de Proximidade	Centralidade de Autovetor	PageRank	Coeficiente de Agrupamento
Maior g.e.	32	2	2495,667	0,011	0,124	10,767	0,015
Maior g.s.	8	11	1298,667	0,009	0,0045	6,079	0,017
Maior coef. agr.	2	0	0	0,006	0,010	0,752	1
Menor g.e.	0	1	0	0,004	0	0,528	0
Menor g.s.	1	0	0	0,004	0	0,554	0
Menor PageRank	0	1	0	0,007	0,018	0,427	0

Tabela 2 do conjunto de interesse AP470