Universidade Federal do Amazonas
Instituto de Computação
Programa de Pós-Graduação em Informática

Guibson Moreira de Souza

Caracterizando Influência e Difusão de Informação no Twitter

Manaus
Junho de 2019

Guibson Moreira de Souza

Caracterizando Influência e Difusão de Informação no Twitter

Trabalho apresentado ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas para obtenção do grau de Mestre em Informática.

Orientador: Prof. Dr. Eduardo Luzeiro Feitosa

Manaus
Junho de 2019

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Souza, Guibson Moreira de
S729c Caracterizando Influência e Difusão de Informação no Twitter / Guibson Moreira de Souza. 2019
85 f.: il. color; 31 cm.

Orientador: Eduardo Luzeiro Feitosa
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Influência do Usuário. 2. Difusão de Informação. 3. Twitter. 4. Interação Social. 5. Rede Social. I. Feitosa, Eduardo Luzeiro II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"Caracterizando Influência e Difusão de Informação no Twitter"

GUIBSON MOREIRA DE SOUZA

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Eduardo Luzeiro Feitosa - PRESIDENTE

Prof. Eduardo James Pereira Souto - MEMBRO INTERNO

Prof. Gilbert Breves Martins - MEMBRO EXTERNO

Manaus, 17 de Junho de 2019

A toda minha família e minha namorada

"Não venci todas as vezes que lutei,
mas perdi todas as vezes que deixei de lutar."

Cecília Meireles

Agradecimentos

Agradeço a Deus por sempre me guiar e me dar sabedoria e saúde.

Agradeço a toda minha família, principalmente aos meus pais Walderilio e Marlete, pois sempre me ensinaram e ainda ensinam muito sobre a vida e além disso sempre se esforçam para eu ter a melhor educação possível, sem vocês eu nada conseguiria, muito obrigado.

Aos meus irmão Gladson e Max, que estão comigo sempre e que sempre me deu conselhos.

A minha namorada Larissa, que sempre confia e nunca duvida no meu potencial, além de me incentivar a chegar cada vez mais longe. Obrigado por sempre estar comigo em todas horas boas ou não, por sempre ser uma amiga e companheira.

Também ao meu orientador Eduardo Feitosa que está comigo desde a graduação e sempre me ajudou nas dificuldades além de ser um excelente profissional da educação.

A todos os membros e amigos do laboratório do ETSS, pelas dicas, contribuições e ótimo convívio durante esses anos.

A todos os amigos externos a Universidade, pelo companheirismo de sempre.

Aos membros da banca examinadora, Eduardo Souto e Gilbert Martins, por aceitarem fazer parte desta importante etapa além das contribuições dadas.

A todo o PPGI (professores e administrativo) pelo ótimo trabalho que desenvolvem para que todos alunos tenham sempre o melhor suporte e a melhor formação acadêmica.

À CAPES pelo apoio financeiro para a realização desta pesquisa.

Resumo

As informações publicadas pelos usuários de redes sociais são, hoje em dia, facilmente utilizadas para analisar tanto o comportamento dos usuários quanto sua capacidade de influênciar e difundir informações dentro da rede social. Embora existam métricas aprimoradas para a identificação do maior número possível de usuários em uma determinada rede, ao analisar cenários específicos percebe-se que tais métricas não conseguem fielmente representar o mundo real. Esta dissertação compartilha da visão que é possível mensurar a influência e a difusão de informação de usuários em redes sociais mesclando características topológicas e baseadas no conteúdo, o que permite avaliar contextos diferentes e empregar características incrementais. Os resultados obtidos, através de bases coletadas dos tópicos de tendência do Twitter, permitem entender que o usuário através de seu comportamento individual e coletivo pode influenciar pessoas e difundir informações, mesmo que este usuário não seja considerado importante ou famoso dentro da rede social analisada, além disso é observado a presença frequente de usuário em determinados nichos.

Palavras-chave: Influência do Usuário, Difusão de Informação, Twitter, Interação Social, Rede Social, Centralidade.

Abstract

The way information is provided by users has become easier to analyze so that their experience is more and more profitable, but it becomes one of the main problems within social networks, which today has several ways to perform interactions. This published information may respond a lot about the user as to their influence or the ability to disseminate information within the social network. Metrics used today are increasingly being improved to identify the maximum users, but when analyzing specific scenarios the metrics can not be fully efficient and faithfully represent the real world. This dissertation shares the view that it is possible to measure the influence and diffusion of information by merging topological and content-based characteristics, since this allows to evaluate different contexts and to use characteristics that can increase others. The results obtained through bases collected from Twitter trending topics allow us to understand that the user through his individual and collective behavior can influence people and spread information, even whether this user is not considered important or famous within the social network analyzed, In addition, the frequent presence of users in certain niches is observed.

Keywords: User Influence, Diffusion of Information, Twitter, Social Interaction, Social Network, Centrality.

Lista de Figuras

1.1 Taxonomia proposta por Al-Garadi et al. [3]. xvi

2.1 Exemplo de um *tweet* xx

2.2 Exemplo de Grafo em Rede Social. xxii

2.3 Grafos rotulados e suas centralidade (A) *in-degree* e (B) *out-degree* xxiv

2.4 Exemplo da medida *Betweenness Centrality* xxv

2.5 Exemplo da medida *Closeness Centrality*. Grafo rotulado A e o cálculo da medida no grafo B. xxvi

4.1 Visão geral da Metodologia proposta xlv

5.1 Visão geral da base 1 relacionando os usuários mais retweetados . liv

5.2 Visão geral da base 1 relacionando os usuários mais retweetados . lvii

5.3 Classificação dos usuários baseados na métrica *Page Rank* lx

5.4 "Usuário#8653" considerado o mais influente utilizando os dois conjuntos de características lxi

5.5 Análise de quantidade *retweets* do "Usuário#8653" por período . . lxii

5.6 Classificação geral dos usuários (Pró e Contra) utilizando a métrica *In-Degree* lxiii

5.7 Classificação dos usuários "Pró" baseados na métrica *In-Degree* . . lxiv

5.8 Classificação dos usuários “Contra” baseados na métrica *In-Degree* lxvi

5.9 Classificação dos usuários “Contra” baseados na métrica *Out-Degree* lxviii

5.10 Análise de quantidade *retweets* do “Usuário#7961” por período . . lxxi

5.11 Análise de quantidade *retweets* do “Usuário#7745” por período . . lxxii

Lista de Tabelas

2.1 Relação dos nó do grafo da Figura 2.5 e sua distância aos demais. xxvi

3.1 Trabalhos relacionados xl

4.1 Lista de métodos da *Search API* xlvi

4.2 Tabela de Bases Coletadas l

5.1 Classificação dos usuários baseados na métrica *in-degree* lv

5.2 Classificação dos usuários baseados na métrica *out-degree* lvii

5.3 Classificação dos usuários baseados na métrica *Page Rank* lix

5.4 Classificação dos usuários "Pró" baseados na métrica *in-degree* . . lxv

5.5 Classificação dos usuários "Contra" baseados na métrica *in-degree* lxvi

5.6 Classificação dos usuários "Pró" baseados na métrica *Page Rank* . lxix

Sumário

1 Introdução **xiv**

1.1 Motivação e Justificativa xiv

1.2 Questão de Pesquisa xvi

1.3 Hipótese xvii

1.4 Objetivos xvii

1.5 Contribuições xvii

1.6 Estrutura do Documento xvii

2 Conceitos Básicos **xix**

2.1 Rede Social xix

2.1.1 Twitter xix

2.2 Influência e Difusão de Informação xxi

2.3 Análise de Redes Sociais xxii

2.3.1 Métricas de Centralidade xxiii

2.4 Discussões xxviii

3 Trabalhos Relacionados **xxix**

3.1 Topologia de Rede xxix

3.2 Baseados em Características do Usuário xxxiii
3.3 Discussões xl

4 Metodologia e Experimentação **xliv**
4.1 Metodologia e Implementação xliv
4.1.1 Etapas da Metodologia xlv
4.2 Bases de Dados l
4.2.1 Racismo l
4.2.2 Política li
4.2.3 Ambiente de Experimentação li
4.3 Discussões lii

5 Resultados **liii**
5.1 Racismo liii
5.1.1 Visão geral liv
5.1.2 Discussão lx
5.2 Política lxii
5.3 Discussões lxxiii

6 Conclusões e Trabalhos Futuros **lxxv**
6.1 Trabalhos Futuros lxxvii

Referências Bibliográficas **lxxviii**

Capítulo 1

Introdução

Com a consolidação da Internet, as pessoas passaram a utilizar redes sociais para encontrar informações sobre assuntos variados, publicar fotos e vídeos, interagir com amigos distantes e compartilhar opiniões. Esse tipo de comportamento tem importância crucial dentro das redes sociais, pois as pessoas são influenciadas e decisões são tomadas com base nas publicações realizadas. Por exemplo: "o que comer?", "o que vestir?" e "o que comprar?" são perguntas que podem ser respondidas através de publicações realizadas por diversos usuários.

Hoje em dia, não é exagero afirmar que as redes sociais tornaram-se a principal fonte de divulgação de conteúdos, auto promoção, venda de produtos, entre outras atividades. A existência desses "nichos" de usuários permitiu que fosse possível perceber a existência de usuários com a habilidade de influenciar outros. Conhecidos hoje como *digital influencers*, estes usuários geralmente possuem um alto poder de convencer outros a utilizarem produtos de uma certa marca ou até mesmo produtos criados por eles mesmos.

Contudo, o mundo dos influenciadores não se resume a produtos. Já foram encontradas contas no Twitter, criadas para: (i) influenciar o andamento de eleições, privilegiando diretamente um candidato [12] e (ii) divulgar rumores, com o intuito de gerar insatisfação e desconfiança dos usuários com os prestadores de serviços da rede social [54]. Para agravar o cenário, o Twitter permite a existência de contas automatizadas, controladas por APIs, que realizam compartilhamentos e publicações, podendo assim difundir informações diversas na rede social.

1.1 Motivação e Justificativa

Com o crescimento do número de usuários nas redes sociais, novas funcionalidades passaram a ser criadas, tais como grupos, comunidades, serviços, e, também, os chamados *market places*. Desta forma, os usuários de redes sociais passaram

a utilizá-las e, assim, obtiveram maior visibilidade, resultando em novos amigos ou seguidores. A partir desse ecossistema, qualquer notícia repercute e se espalha pela rede social em questão de minutos, algo antes alcançado apenas por usuários considerados famosos. Desta forma, é cada vez mais frequente usuários comuns conseguirem movimentar uma rede social com temas relevantes (ou não) e conseguirem obter certo grau de influência, às vezes até mesmo em cenário mundial.

Tomando o Twitter como exemplo, a existência de usuários influentes contribui para uma melhor experiência de outros usuários perante à rede social. Por ser uma plataforma de mensagens consideradas curtas (280 caracteres), onde podem ser anexados arquivos, as publicações tornam-se geradoras de *retweets*, o que acarreta uma difusão do conteúdo para seus seguidores e outros usuários que não necessariamente são seguidores do publicador da mensagem original.

Identificar estes tipos de usuários influentes tem ganhado importância nos últimos anos. Normalmente, empresas têm utilizado esses usuários para divulgarem seus produtos, por acreditarem ser uma forma mais barata de conseguir que sua marca tenha um amplo alcance dentro da rede social [43]. Obviamente, usuários influentes também possam ser usados em objetivos menos nobre como espalhar conteúdos indesejáveis dentro da rede (vírus, spam, rumores e até difamação) [46]. A questão é que o processo de identificação de usuários influentes é tratado de diferentes maneiras, geralmente porque as redes sociais tem diferentes finalidades.

Al-Garadi et al. [3] apresentam três formas de identificar usuários influentes: (1) **redes sociais**, onde são descritas as conexões existentes entre os usuários. No Twitter, por exemplo, se o usuário A segue o usuário B existe uma conexão entre eles; (2) **redes de propagação**, que são caracterizadas pela forma como as informações são espalhadas ou propagadas de um usuário a outro, existindo ou não uma conexão entre eles. Por exemplo, no Twitter é caracterizado através dos *retweets*; (3) **redes de interação**, que descrevem a capacidade de um usuário envolver outros em uma conversa. No Twitter é representado através de *tags* ou menções.

Contudo, avaliar diferentes tipos de redes pode não ser eficaz quando não se está direcionando pra quem ou qual evento está se querendo avaliar. Em outras palavras, escolher um nicho para avaliação em uma propagação de certo conteúdo é mais eficiente na identificação de usuários influentes do que avaliar uma rede social como um todo, dado que um usuário influente em um nicho específico pode não ser facilmente identificado em outro nicho ou não ter a mesma facilidade em difundir suas publicações.

Al-Garadi et al. [3] propuseram uma taxonomia (Figura 1.1) que caracteriza os diferentes tipos de influenciadores encontrados em redes sociais. Os autores

afirmaram que analisar influência de usuários é essencial para entender a rápida adoção de tendências específicas. Com isso, relacionaram temas específicos que consideraram importantes, pois para alguns deles são criados produtos finais, tais como, *marketing* [57], saúde [56] e análise de produtos [55].

Figura 1.1: Taxonomia proposta por Al-Garadi et al. [3].

Apesar da taxonomia, investigar a influência de um usuário dentro de uma rede social é um desáfio. Os problemas mais identificados na literatura sobre o assunto são: (1) encontrar métricas adequadas que possam identificar estes tipos de usuários em diferentes contextos; (2) como avaliar se um usuário sempre continua com status de influente no mesmo contexto e (3) como os usuários comuns se comportam perante a difusão de informação na rede social. Para isso foram propostas hipóteses que possam comprovar estes problemas caracterizados.

1.2 Questão de Pesquisa

A partir da contextualização e motivação apresentadas, esta dissertação tem como objetivo investigar e responder às seguintes questões:

- **QP1** "É possível identificar usuários influentes em nichos ou temas específicos em redes sociais?"
- **QP2** "As métricas encontradas hoje na literatura conseguem representar fielmente influência e difusão de informação em redes sociais?"

1.3 Hipótese

A hipótese definida para esta dissertação considera o seguinte cenário:

- A aplicação das métricas existentes e comumente empregadas na identificação de usuários influentes no Twitter, produz o mesmo resultado quando aplicadas a nichos ou temas específicos.

1.4 Objetivos

O objetivo central desta dissertação é definir e demonstrar que as atuais métricas para identificação de usuários influentes não são aplicáveis em qualquer cenário ou evento em redes sociais, através da seleção e avaliação de características e aplicação de algoritmos consolidados da literatura. A finalidade é auxiliar pesquisadores ou terceiros na identificação de usuários influentes em áreas específicas.

Como objetivo específico, pretende-se definir primeiramente quais características são baseadas na estrutura da rede social e centrada ao usuário, com isso aplicar diferentes conceitos de influência e difusão de informação dentro do cenário ou evento coletado.

Além deste, outro objetivo específico é a criação de um script que analisa especificamente o usuário através de sua rede de contatos, com a finalidade de poder analisar mais detalhadamente sua influência e sua capacidade de difundir informação.

1.5 Contribuições

As principais contribuições esperadas desta dissertação são:

- Uma metodologia que engloba fase de coleta e avaliação das características e algoritmos, criando novas etapas que avaliam para diferentes tipos de redes.

- A identificação de um conjunto de características que permitem diferenciar usuários influentes de usuários não influentes, e também os usuários difusores de informação.

1.6 Estrutura do Documento

Este documento está organizado em 5 capítulos. No Capítulo 2 são apresentados os conceitos básicos necessários para a compreensão desta proposta. O Capítulo

3 expõe os trabalhos relacionados encontrados na literatura. O capítulo 4 detalha a solução proposta, apresentando uma metodologia para detecção de contas maliciosas. Por fim, no capítulo 5 são apresentadas as atividades atuais e futuras.

Capítulo 2

Conceitos Básicos

Este Capítulo apresenta os conceitos básicos utilizados neste trabalho, os quais são necessários para o desenvolvimento e entendimento sobre os temas abordados.

2.1 Rede Social

Uma Rede Social é definida por Ellison et al. [20] como um serviço que permite indivíduos construírem perfis públicos ou semi-públicos, gerenciar uma lista de amigos com quem compartilham informações e percorrer a lista de outros usuários cadastrados na rede. Já Wives [49] define rede social como um conjunto de pessoas ou grupos onde existe algum tipo de relação entre si. Portanto, assim como em redes de computadores, onde máquinas são interligadas, redes sociais possuem o mesmo aspecto, já que conectam pessoas formando as redes [45].

Cada rede social possui uma determinada finalidade. Algumas visam apenas um grupo social enquanto outras abrangem um número maior de usuários. O LinkedIn[1], por exemplo, foi desenvolvido com o propósito voltado para negócios, enquanto o Facebook[2] foi criado com o propósito de usuários compartilharem informações e experiências. Já o Twitter foca no compartilhamento de notícias, fotos, vídeos e seus usuários frequentemente a utilizam para atualizar o que estão fazendo.

2.1.1 Twitter

É uma rede social bastante conhecida e utilizada em todo mundo [42]. Criada em 21 de Março de 2006, permite aos usuários enviarem e receberem atualizações de outros contatos em texto de até 280 caracteres, incluindo *emojis*, fotos, vídeos,

[1] www.linkedin.com
[2] www.facebook.com

URLs e menções a outros usuário. As publicações são conhecidas como *tweets* (Figura 2.1).

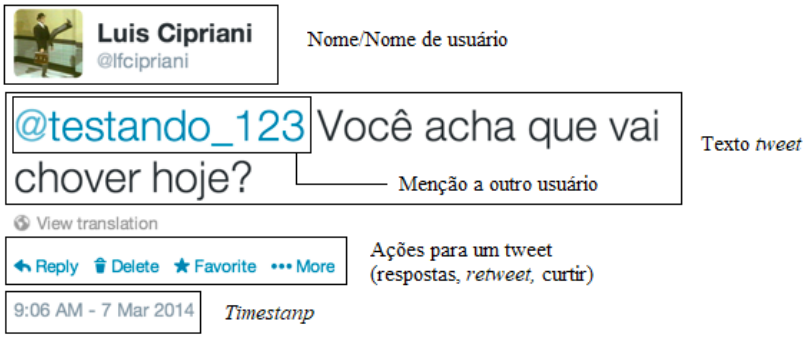

Figura 2.1: Exemplo de um *tweet*

Os principais termos utilizados por usuários no Twitter são:

- *Tweets*: Uma mensagem no Twitter contendo no máximo 280 caracteres, incluindo *emojis*, fotos e *URLs*.
- Seguindo e Seguidores: Seguidores são os usuários que estão seguindo um usuário específico e seguindo são os usuários que o usuário segue.
- *Retweets*: Um *tweet* que foi compartilhado novamente com todos os seguidores de um usuário.
- *Hashtags*: É usado para marcar palavras-chave ou tópicos em um *tweet* para torná-lo facilmente identificável para fins de pesquisa, caracterizado por "#".
- Menções: *Tweets* podem incluir respostas e menções para outros usuários, precedendo seus nomes de usuário com sinal "@".
- Tópicos de Tendência (*Trending Topics*): uma lista em tempo real das palavras mais postadas no Twitter(que acompanham "#") em todo o mundo.

No Twitter, as contas podem ser públicas ou privadas. As contas públicas permitem que todo o histórico de publicações de um usuário possa ser visto por outros sem necessariamente estar seguindo a pessoa. Já as contas privadas permitem apenas que os seguidores possam ter acesso as suas publicações, curtidas e *retweets*. As formas de relacionamentos entre usuários no Twitter são: usuário para usuário, *tweet* para *tweet*, *tweet* para usuário e usuário para *tweet*. Modelando essa formas de relacionamento com grafos é possível perceber que podem

existir de diversas maneiras: menções, *retweets*, seguir/seguidor e respostas (*replies*). Com isso, pesquisadores tem utilizado estes tipos de relacionamentos para mensurar influência dentro do Twitter. O tipo de relacionamento mais utilizado é o *retweet*, pois pode representar tanto a questão de influência de usuário quanto serve para analisar o conceito de difusão de informação dentro da rede.

Para obter estes dados do Twitter e poder analisá-los, é disponibilizada uma API[3] para desenvolvedores, onde estão acessíveis um conjunto de funções, protocolos e ferramentas usadas para construir aplicações que se comunicam com os serviços fornecidos. A chamada API *REST*, que é utilizada neste trabalho, permite ler e escrever dados no Twitter através de métodos *GET* e *POST*, e a sua resposta é apresentada em um formato JSON.

2.2 Influência e Difusão de Informação

De acordo com Ferreira [22], influência significa ato ou efeito de influir, ação que uma pessoa ou coisa exerce sobre outra, poder, crédito, prestígio. Portanto, trazendo para o contexto de redes sociais, influência é geralmente atribuída a capacidade de um usuário manipular ou alterar sentimentos, atitudes ou comportamentos de outros usuários em uma rede [19]. Logo, a influência que um usuário obtém dentro da rede é oriunda da chamada difusão de informação, ou seja, o espalhamento dessas informações por meio de funcionalidades que a rede social fornece (curtir, compartilhar e/ou amizades, por exemplo).

Diversos autores propuseram formas de definir a influência dentro da rede social, pois acreditavam que dependendo da disposição dos usuários ou do evento que está se avaliando os usuários se comportavam de maneiras diferentes. Assim, criaram termos para definir usuários influentes como, por exemplo, líderes e inovadores [14], engajadores [16] e difusores de informação [30].

Riquelme and González-Cantergiani [40] conduziram uma revisão da literatura onde exploraram os diversos tipos de métricas que podem caracterizar usuários influentes. Para isso, criaram taxonomias (atividade, popularidade e medidas de influência) para ajudar a computar diferentes características dos usuários. Essas taxonomias levam em conta características que envolvem o usuário e podem ser coletadas diretamente do Twitter e em algoritmos baseados na topologia da rede.

As diversas formas como usuários influentes são identificados é algo bastante estudado e que ainda possui lacunas a serem investigadas, como será mais detalhado no Capítulo 3. Inicialmente, usuários influentes eram identificados através do número de pessoas que seguiam e os seguidores que tinham [3]. Entretanto,

[3]https://dev.twitter.com

hoje é possível que um usuário comum seja ou se torne influente dentro da rede, pois dependendo do alcance da publicação irá ganhar novos seguidores e seu conteúdo será cada vez mais espalhado dentro da rede social.

2.3 Análise de Redes Sociais

Usuários de Redes Sociais vivem em um mundo interligado através de interações sociais, tais como, compartilhamento de informações, mensagens ou opiniões. A representação destes usuários é realizada, em grande parte das vezes, na forma de grafos, onde cada usuário é um nó (vértice) e suas interações na rede são arestas [51]. A Figura 2.2 exemplifica essa representação.

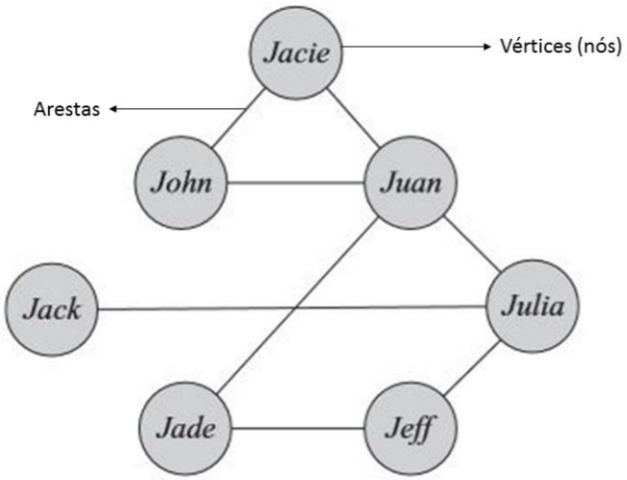

Figura 2.2: Exemplo de Grafo em Rede Social.

Assim, usuários em redes sociais podem ser representados como um grafo $G = (V,E)$, onde V é o conjunto de nós e E é o conjunto de arestas que representa como o nós estão relacionados. Transformando esses conceitos no linguajar utilizado no Twitter, V é considerado o conjunto de usuários e E as relações (interações) entre os usuários, que podem acontecer através de *retweets*, menções, comentários, entre outros.

Como neste trabalho será analisado a rede social no aspecto de difusão de informações e influência, existem algoritmos capazes de mensurar nós influentes para diferentes tipos de rede. Essa importância de um nó em uma rede é calculada usando métricas importadas de teoria dos grafos [34][23].

2.3.1 Métricas de Centralidade

Uma vez que é possível empregar grafos para representar usuários em redes sociais, surge a questão de como averiguar o comportamento ou interação desses usuários na rede. Como responder a questões do tipo: "Qual é o principal usuário dentro de uma rede?", "Quais padrões de interação são comuns entre usuários?". Existem diferentes métricas na literatura para tal propósito.

De acordo com Beveridge and Shan [7], as métricas mais utilizadas são aquelas baseadas em centralidade, uma vez que definem o nó mais importante dentro de uma rede. Como esta pesquisa precisará aplicar esse tipo de métricas, serão apresentadas cinco (05) medidas relacionadas a centralidade mais empregadas na literatura.

Degree Centrality

Degree Centrality considera o número de arestas incidentes que um vértice possui, ou seja, o número de ligações que certo nó possui. A medida *degree centrality* é altamente eficaz para mensurar a influência ou importância de um nó, podendo, por exemplo, ser empregada para medir tudo o que circula dentro de uma rede e passa por determinado nó [24]. Para calcular a *degree centrality* de um único nó v, para o grafo $G = (V,E)$ com $|V|$ vértices e $|E|$ arestas, utiliza-se a equação 2.1.

$$C_D(v) = deg(v) \tag{2.1}$$

Entretanto, se o interesse é calcular a centralidade de um grafo levando em consideração todos os seus vértices, utiliza-se a equação 2.2.

$$C_D(G) = \frac{\sum_{i=1}^{|V|}[C_D(v*) - C_D(v_i)]}{\sum_{j=1}^{|V|}[C_D(v*) - C_D(v_j)]} \tag{2.2}$$

De acordo com Freeman [24], *degree centrality* permite medir a contagem do número de arestas direcionadas ao nó, chamada de *in-degree*, bem como medir a contagem de aresta que o nó encaminha aos outros, chamada de *out-degree*. A Figura 2.3 representa essas duas medidas - grafo A para a medida *in-degree* e o grafo B para a medida *out-degree*, onde percebe-se na claramente uma variação nos valores dos nós.

Ainda com base no trabalho de Freeman [24], as medidas de *in-degree* e *out-degree* só podem ser obtidas em grafos conectados. Desta forma, quando as arestas estão associadas à aspectos positivos, como amizade ou colaboração, a

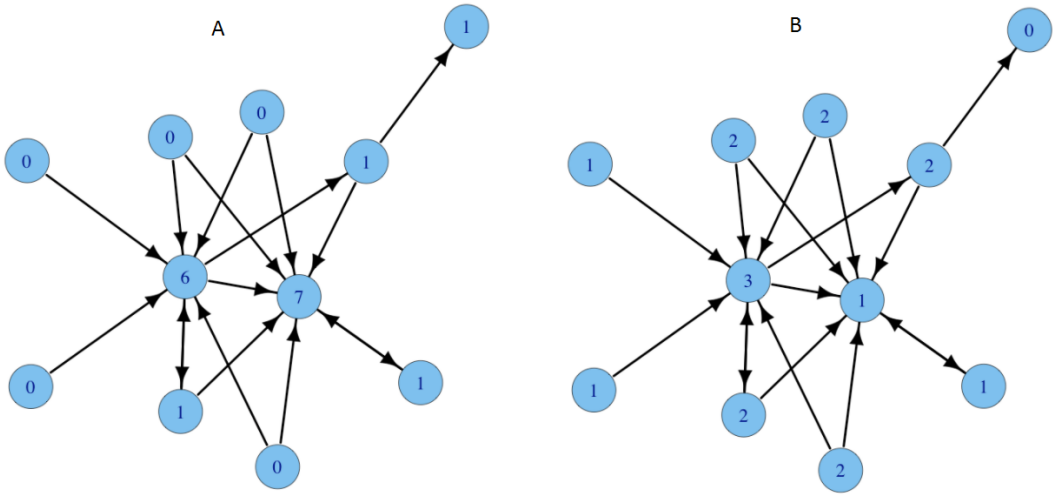

Figura 2.3: Grafos rotulados e suas centralidade (A) *in-degree* e (B) *out-degree*

medida *in-degree* pode ser interpretada como uma forma de popularidade enquanto a medida *out-degree* como uma forma de sociabilidade.

Betweenness Centrality

Betweenness Centrality é uma medida de centralidade que no decorrer do seu cálculo leva em consideração os caminhos mínimos [10]. Para cada par de vértices em um grafo conectado, existe pelo menos um caminho mais curto entre os vértices. Matematicamente, pode-se obter o *betweenness* v_i de um nó i através da equação 2.3, onde σ_{st} é o valor total de caminhos mínimos entre os vértices s e t, e $\sigma_{st}(i)$ é a quantidade de caminhos mínimos que passam pelo nó i.

$$v_i = \sum_{s \neq t \neq z} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{2.3}$$

Um exemplo de aplicação desta métrica é em redes de telecomunicações, onde um nó com maior valor de *betweenness centrality* tem mais controle sobre a rede, porque mais informações passam através dele. A Figura 2.4 exemplifica o uso de *Betweenness Centrality*.

Percebe-se na Figura 2.4 que o vértice 4 tem o maior valor, pois ele atua como uma espécie de "ponte" entre o lado esquerdo e direito do grafo.

Closeness Centrality

Closeness Centrality é a medida que calcula a distância de um nó para todos os vértices da rede, como forma de medir a importância deste nó. Em grafos

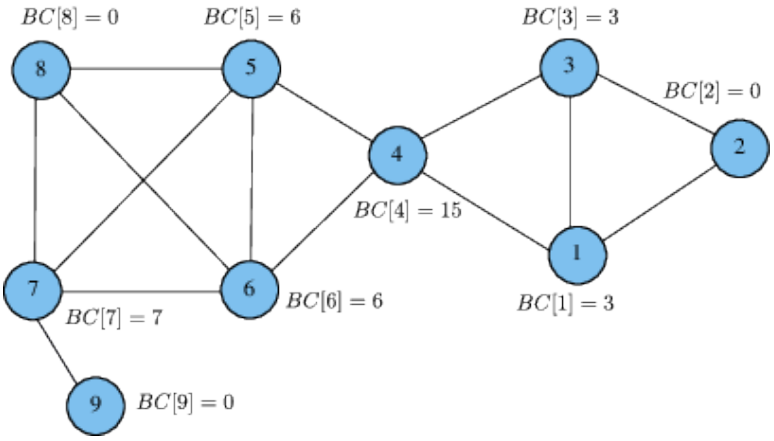

Figura 2.4: Exemplo da medida *Betweenness Centrality*

conectados existe uma distância média entre todos os pares de nós, definida pelo comprimento de seus caminhos mais curtos [41]. Em uma rede social, por exemplo, uma pessoa com menor distância média para os outros usuários (vértices) pode achar que suas opiniões atingem outros usuários (vértices) na comunidade mais rapidamente do que a opinião de alguém com maior distância média.

Matematicamente os valores gerados nos vértices do grafo na Figura 2.5 podem ser obtidos através da equação 2.4, com $v \in V$, $n = |V|$ e dist(x,y) sendo a distância entre o nó x e o nó y.

$$C(v) = \frac{n-1}{\sum_y dist(y,x)} \tag{2.4}$$

Closeness centrality não é uma medida de centralidade igual no sentido das medidas anteriores, uma vez que dá valores baixos para nós mais centrais e valores altos para os menos centrais, exatamente o oposto das outras medidas de centralidades.

Assim como foi definido anteriormente, esta medida de centralidade obtém a distância de um nó para toda sua rede, através da soma de todos os caminhos de pares de nós. A Figura 2.5 representa um grafo onde os nós estão rotulados (1 a 5) - a esquerda - e o mesmo grafo com o cálculo da *closeness centrality* de acordo com a Equação 2.4. A Tabela 2.1 relaciona a soma dos caminhos mais curtos entre os nós da rede.

Eigenvector Centrality

Eigenvector Centrality é uma extensão natural da *degree centrality*. Esta medida de centralidade assume que um vértice importante está conectado com outros nós

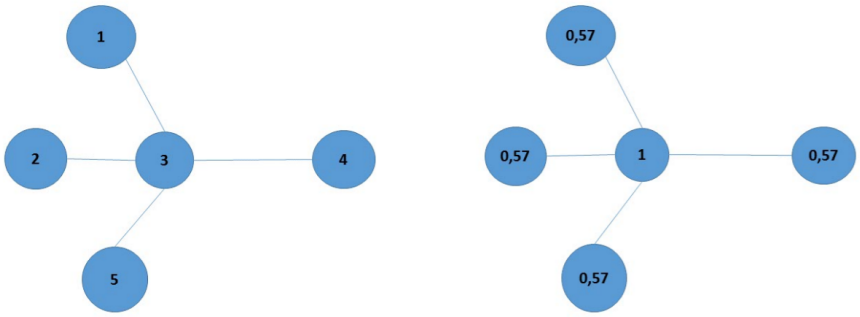

Figura 2.5: Exemplo da medida *Closeness Centrality*. Grafo rotulado A e o cálculo da medida no grafo B.

Tabela 2.1: Relação dos nó do grafo da Figura 2.5 e sua distância aos demais.

Nós	**1**	**2**	**3**	**4**	**5**	$\sum_y dist(y,x)$	$C(v) = \frac{n-1}{\sum_y dist(y,x)}$
1	0	2	1	2	2	7	0,57
2	2	0	1	2	2	7	0,57
3	1	1	0	1	1	4	1
4	2	2	1	0	2	7	0,57
5	2	2	1	2	0	7	0,57

importantes, onde a importância x_v de um vértice i pode ser medida pelo somatório da importância dos seus vizinhos, pois eles atribuem pontuações relativas a todos os nós da rede.

Matematicamente, pode-se obter o *eigenvector centrality* aplicando-se a equação 2.5, onde as arestas para os nós de alta pontuação contribuem para a pontuação do nó em questão. Para acompanhar a importância destes vizinhos é utilizado a matriz de adjacência de um grafo.

$$v_i = \frac{1}{\lambda} \sum_k a_{k,i} v_k \tag{2.5}$$

Dado uma matriz de vértices adjacentes A com entradas $a_{i,j}$ e $\lambda \neq 0$ sendo um autovalor da matriz A, este cálculo é aplicado sobre cada vértice da rede, não sendo necessário ter várias arestas conectadas a um vértice para se obter um alto *eigenvector* [36].

Page Rank Centrality

Page Rank Centrality é uma medida de centralidade que assume que a contribuição de centralidade dos vértices não é a mesma e deve ser penalizada proporci-

onalmente de acordo com a quantidade de vizinhos, ou seja, se um vértice com alto grau possuir muitos vizinhos para dividir sua importância, sua centralidade será mais penalizada que os outros nós com menos vizinhos.

Existem três (03) fatores distintos que determinam o *page rank* de um nó: (i) o número de links que recebe, (ii) o link de propensão das arestas e (iii) a centralidade das arestas.

O *page rank* de um nó é calculado conforme a equação 2.6, onde $a_{k,i}$ é a matriz de vértices adjacentes do nó a qual se aplica a equação, d_d é o grau de saída do vértice k, α e β são constantes e v_k é o valor de *pagerank* do vértice vizinho [11].

$$v_i = \alpha \sum_k \frac{a_{k,i}}{d_k} v_k + \beta \tag{2.6}$$

Algoritmos Relacionados

Existem outras métricas de centralidade que podem ser relacionadas para analisar redes sociais no contexto de redes complexas, cada uma delas possuindo finalidades diferentes e havendo também suas complexidades. Outra métrica baseada em caminho mais curto é a centralidade de *Katz*. Esta métrica calcula a influência de usuários levando em consideração todos os caminhos de rede. Sendo assim, a influência é determinada através de todas as arestas que passam pelo nó [25]. A sua principal diferença para a métrica *Closeness* é que *Katz* considera apenas os comprimentos de caminho curto [32]. Outra diferença que a centralidade de *Katz* possui em relação a *Closeness* é atribuir uma certa pontuação mínima a cada usuário na rede. Apesar de possuir uma boa suposição matemática para analisar redes e identificar nós importantes, a métrica possui um ponto negativo que é a sua complexidade, que limita uma analisa em grandes redes [32].

Outra métrica também relacionada por estudos que analisam interações com a finalidade de encontrar influenciadores e difusão de informações entre usuários é o *k-core*, também conhecido como *k-shell*, onde pressupõe que a localização dos usuários é mais importante do que a conexão direta para o cálculo de difusão de influência [26]. A principal vantagem desta métrica é que mesmo em redes com dados incompletos ela é capaz de calcular influência mais eficientemente que outras abordagens. Sua desvantagem é que sua saída possui dois conjuntos de verdadeiros nós influentes: os que possuem um nível de k, que reflete corretamente a sua influência; e os com alto k, mas não são considerados difusores de influência [31].

2.4 Discussões

Neste Capítulo foram apresentados assuntos relacionados que levaram ao desenvolvimento desta dissertação, tais como, Redes Sociais, Twitter, Grafos (Redes Complexas) para análise de redes sociais e métricas existentes. Esses assuntos são importantes pois cada um deles podem prover informações em forma de dados que podem resolver análises sobre qualquer evento ou rede específica.

Apesar de existirem diversas redes sociais e inúmeras métricas para realizarem essa avaliação, ainda existem algumas lacunas encontradas, por exemplo, a complexidade de avaliar grandes redes, identificar e saber diferenciar o que são influenciadores e difusores de informação e os impactos que estes usuários podem gerar. O Capítulo 3 relaciona os trabalhos existentes da literatura identificando pontos positivos e negativos de cada abordagem proposta.

Capítulo 3

Trabalhos Relacionados

Neste Capítulo são apresentados os estudos e trabalhos que identificam ou caracterizam usuários influentes em redes sociais. São apresentadas métricas convencionais bem como aquelas propostas em outras áreas, mas que foram combinadas às redes sociais, permitindo a criação de novos conceitos e medições. Para tanto, este Capítulo é organizado em duas seções que representam duas diferentes formas de identificar usuários influentes, sejam eles difusores de informação ou somente usuários populares. Ao final são discutidos os pontos fortes e fracos das abordagens apresentadas.

3.1 Topologia de Rede

Nesta seção, os trabalhos discutidos são relacionados à atributos que envolvem os usuários em uma análise de nível global na rede através das interações já conhecidas como, por exemplo, *retweets* e/ou menções. Os trabalhos apresentam distintas formas de avaliar influência e difusão de informação dos usuários, analisando seu comportamento seja por meio de Teoria dos Grafos ou pelo conceito de comunidades.

Dubois and Gaffney [18] realizaram uma análise comparando cinco (05) métricas (*Indegree*, *Eigenvector*, *Clustering*, Conhecimento e Interação) comumente usadas para identificar usuários influentes em redes sociais em eventos relacionados a política. O estudo baseia-se em interpretações de definições de influência para responder a questão de pesquisa "Quais políticos são os mais influentes dentro das duas maiores comunidades políticas canadense?".

A metodologia utilizada para o trabalho foi coletar todos os *tweets*, durante um período de duas semanas, que possuíam ao menos uma das duas *hashtags* (#CPC e #NDP) que representam partidos políticos. Após a coleta, um grafo de relacionamento baseado em seguidores foi criado, aplicando-se métricas de análise

de rede (*Indegree*, *Eigenvector*, *Clustering*) e análise de conteúdo (Conhecimento e Interação). Como forma de avaliar e comparar a classificação dos usuários em relação as métricas, os autores utilizaram o coeficiente de correlação *Kendall's* τ - que para um valor τ alto as métricas concordam e para um valor τ baixo discordam ou divergem para um τ igual a zero.

Como resultado, a classificação após o coeficiente ser aplicado mostrou um tendência similar entre as métricas *Indegree* e *Eigenvector* para as duas *hashtags* coletadas. Segundo os autores, essas duas métricas são importantes para identificar políticos tradicionais como influentes. Por outro lado, as métricas que envolvem interações e análise do conteúdo identificaram diferentes grupos de influenciadores, como comentaristas políticos e blogueiros.

Räbiger and Spiliopoulou [39] propuseram um *framework* modular para identificação de usuários influentes e não influentes em uma rede social. Dividido em duas partes, o primeiro segmento do *framework*, chamado *InfluenceLearner*, extrai um grafo de relacionamento e um grafo de interação social, calcula as propriedades de rede entre eles e em seguida utiliza aprendizagem supervisionada para classificar os usuários. A decisão se um usuário é influente ou não é binária. A segunda parte do *framework*, chamado *SNAnnotator*, coleta os dados da rede social (usuários e *tweets*) e os rotula de forma manual, preparando-os para a extração de atributos.

Os autores utilizaram uma abordagem proposta por Bigonha et al. [8] ao criar dois tipos de grafos: um de relação (conceito de amizade) e um de interação (*retweet*, menções e respostas). Para que o componente *InfluenceLearner* aprenda os atributos que caracterizam influência, foi processada a rede social original e dela foi derivada um conjunto de atributos divididos em quatro (04) classes: (i) **estrutura da comunidade**, que englobam comunidades conectadas, arestas entre comunidades e arestas entre comunidade; (ii) **atividade dos usuários com seus seguidores**, ou seja, interações e *tweets*; (iii) **qualidade do conteúdo publicado pelos usuários**, mensurado pela taxa de seguindo/seguidores e qualidade do *tweet*; e (iv) **conceitos de centralidade** *in-degree*, *closeness* e *eigenvector*.

Os resultados mostraram que existem atributos (pertencentes as quatro classes) preditivos de influência que são associados ao nível de atividades do usuário e seu envolvimento nas comunidades. Os autores afirmam que o grafo de relação permite uma melhor identificação de usuários influentes em comparação com o grafo de interação, pois as relações dos usuários incorporam mais dados que podem ser usados pela aprendizagem supervisionada. Por último, as diferenças entre um usuário influente e não influente são pequenas, não existindo um único atributo capaz de separá-los.

Al-Garadi et al. [2] investigaram as propriedades de diferentes topologias es-

truturais da rede de um usuário influente que propaga informações em redes sociais. Para tanto, propuseram uma representação topológica para redes sociais, levando em conta as interações de multicamadas, ou seja, analisando individualmente em grafos as relações entre usuários (*retweets*, seguidor e menções) e ao final criando um grafo englobando todos os grafos analisados. Outra finalidade do trabalho foi avaliar a representação topológica quanto a sobreposição de arestas como peso. Este cálculo é realizado através da soma dos três tipos de interações na utilizando o grafo com multicamadas.

Os autores abordaram questões que não haviam sido respondidas por trabalhos anteriores, tais como "Diferentes representações de redes topológicas interferem no desempenho das métricas usuais de identificação de propagadores influentes em uma rede social?" e "Qual a melhor representação topológica da rede para identificar precisamente influenciadores no contexto de redes sociais?".

Foram utilizadas duas bases reais, obtidas do Twitter, e três métricas (*degree centrality*, *Page Rank* e *k-core*) para avaliar o desempenho na identificação de usuários influentes em diferentes topologias de rede. A eficácia destas métricas foi avaliada através da comparação da lista de classificação obtida por cada métrica com a lista de classificação obtida pelo rastreamento de arestas de difusão na dinâmica real de informação [38].

Os resultados mostraram que melhorias na precisão da identificação de usuários influentes não se baseiam apenas no aprimoramento de métricas de identificação, mas também no desenvolvimento de uma topologia de rede que represente a difusão da informação. Além disso, os resultados mostraram que não há uma métrica para identificação de usuários influentes que sempre seja excelente em qualquer das topologias de rede representadas no trabalho. Portanto é necessário entender como o conjunto de dados da rede é extraído e como os usuários dentro da rede estão interagindo para identificar os melhores algoritmos possíveis.

Em Al-garadi et al. [1], os autores aperfeiçoaram o método *k-core* [47] para redes sociais, propondo um novo método de ponderação de arestas baseado na interação entre os usuários. A ideia dos autores foi mostrar que a ponderação de arestas é um fator importante na quantificação da capacidade de disseminação do usuário em redes sociais, ou seja, uma forte interação entre os usuários pode resultar em uma alta probabilidade de espalhar informações na rede enquanto uma fraca interação entre usuários pode indicar uma baixa probabilidade de disseminação de informação na rede.

Para tanto, utilizaram interações entre usuários do Twitter a partir de *retweets*, já que uma rede formada por este tipo de interação é melhor em descrever a propagação de conteúdo [15]. Em seguida, a aresta entre os usuários foi ponderada usando o número de interação entre usuários extraídos de informações do *retweet* e menções. Como forma de avaliação do método foram utilizadas duas

bases: a primeira utilizada em [15] e a segunda é uma rede de relacionamento recíproco do Twitter utilizada por [48]. A eficiência do método foi validada calculando suas funções de imprecisão em comparação com outras métricas, tais como, *PageRank*, *degree* e *k-core* original [26]. Além desta forma de calcular a eficiência do método, os autores utilizaram a taxa de reconhecimento para analisar o desempenho de cada métrica na identificação dos usuários influentes. O método proposto alcançou melhor desempenho para ambas, concluindo ter um desempenho superior tanto na questão de identificação de usuários influentes quanto na difusão de informação pelos usuários.

Zhuang et al. [58] propuseram um método chamado *SIRank* que não utiliza as métricas de centralidade já conhecidas na literatura, mas realiza uma análise da rede dos usuário baseando-se em características locais, utilizando-as para verificar seu comportamento dentro da rede. De acordo com os autores, as características utilizadas, como por exemplo *retweets* e posição do usuário dentro da rede, são úteis para realizar a análise da difusão de informações e a influência dos usuários. Vale destacar que o formato cascata utilizado como posição do usuário na rede é similar ao conceito utilizado no *Page Rank*.

Para validar o método os autores realizaram comparações com outras métricas derivadas do *Page Rank*, tais como, *Tunk Rank* e *Retweet Rank*, analisando-as nos aspectos de cobertura e predição. *SIRank* obteve uma boa cobertura, próxima aos resultados do *Page Rank* e sendo também o melhor método dentre os comparados na métrica de predição. Logo, o *SIRank* pode possuir uma eficácia para medir usuários difusores e influentes em redes sociais.

Alp and Öğüdücü [5] propuseram uma nova metodologia, chamada *Personalized PageRank*, que integra informações coletadas da topologia do usuário e das suas atividades no Twitter. A metodologia tem o objetivo de determinar os usuários influenciadores de tópicos e aqueles especialistas em um determinado evento. As informações relacionadas as atividades do usuário são: taxa de foco, atividade, autenticidade e velocidade de obter reação.

Foram selecionados 20 usuários em diferentes tópicos (política, esporte, TV e religião). Foram coletados seus usuários e seguidores de uma maneira ampla, até que um número suficiente de usuários fosse obtido. Após a coleta de dados dos usuários o passo seguinte foi pré processar a base e identificar os tópicos ou eventos dos *tweets* para remover informações desnecessárias para futuras análises. Dos usuários coletados foi possível identificar parâmetros, como por exemplo, *tweets* de um usuário, *tweets* de um usuário específico publicados em um tópico específico e quantidade de *retweets* por usuário. Estes parâmetros puderam ser calculados, pois são importantes no processo de identificar usuários influentes em específico. Com esses parâmetros foi possível calcular as características relacionados pelos autores.

O resultado, em um grande conjunto de dados, mostrou que o uso de recursos específicos do usuário em tópicos particulares afeta positivamente a identificação de influenciadores e leva a maior difusão de informações. Os autores concluíram que a divisão da rede global para uma rede que represente tópicos é uma abordagem eficaz para determinar usuários influentes e consideram útil incorporar recursos do usuário aos recursos da rede. Porém, o desempenho de vários recursos específicos do usuário depende da topologia da rede.

Yang et al. [50] desenvolveram um método para identificação de usuários influentes, avaliando diferentes tipos de interações na rede social. Os autores usaram conceitos relacionados às métricas de centralidade *closeness* e *betweenness*. O método proposto modifica a originalidade da métrica *betweenness*, calculando o grau do *closeness* entre os nós e, em seguida, criando pesos arestas com esse cálculo de *closeness*.

Foram construídas pequenas redes para realização de experimentos onde foram definidos os usuários influentes como os nós iniciais na rede. Os autores avaliaram o valor máximo de nós infectados por eles na rede usando o modelo Cascata Independente, pois pelo efeito da difusão a eficácia dos usuários poderia ser determinada. Os resultados demostraram que os usuários influentes eram semelhantes. Porém, após selecionar a parte diferente para outros experimentos observou-se que o efeito de difusão dos nós é melhor no algoritmo *closeness* do que o *betweenness*.

3.2 Baseados em Características do Usuário

Nesta seção os trabalhos discutidos são relacionados a atributos que podem ser extraídos do usuário através de requisições à API do Twitter, tais como, nome de usuário, *retweets*, curtidas, *tweets*, seguidores, entre outros. Os trabalhos apresentam diferentes formas de avaliar influência e difusão de informação dos usuários, implementando técnicas ou agrupando atributos já conhecidos.

Cha et al. [13] utilizaram dados coletados do Twitter e apresentaram uma comparação entre três medidas de influência: *in-degree*, *retweets* e menções. Assim, investigaram a dinâmica do usuário em tópicos e horários. Para isso, coletaram um conjunto de dados de 54 milhões de usuários que produziram mais de 1,7 bilhões de *tweets*, com mais de 2 bilhões de arestas. Dentre esses usuários, foram ignorados aqueles que publicaram menos de dez *tweets* durante a coleta, resultando em mais de 6 milhões usuários. Porém, os usuários ignorados também foram analisados para investigar como eles interagiam com os usuários ativos.

As medidas de influência dos usuários ativos foram calculadas e comparadas entre si. Para não comparar os valores obtidos de forma direta, os autores utilizaram a ordem relativa das classificações dos usuários como uma medida de

diferença. Assim, os usuários ativos foram classificados por cada medida, de modo que a classificação quanto mais próximo de 1 indica o usuário mais influente. Logo após a aplicação das medidas, os autores quantificaram como a classificação dos usuários varia em diferentes medidas e examinaram quais tipos de usuários tem a melhor classificação para uma determinada medida. Para isso, utilizaram o coeficiente de correlação de Spearman's como forma de medir a força da associação entre dois conjuntos de classificação.

Após a classificação nas três medidas, os autores visitaram as páginas do Twitter de cada um dos 20 usuários melhor classificados de cada medida. Os usuários mais seguidos eram desde figuras públicas até fonte de notícias (CNN, New York Times e Barack Obama), concluindo que a medida *in-degree* é útil quando se quer identificar usuários que recebem muita atenção do público através de interações individuais. Os mais retweetados foram serviços de conteúdo (TweetMeme) e empresários (Guy Kawaki, por exemplo). Diferente da medida *in-degree*, os *retweets* representam a influência de um usuário além do domínio de interação um para um, pois *tweets* considerados populares podem se propagar através da rede sem precisar ser de conhecimento da rede local do usuário. De acordo com os autores, os mais mencionados foram usuários considerados celebridades. Usuários comuns normalmente mostram uma grande admiração por celebridades, logo publicam regularmente *tweets* que mencionam eles, mas não necessariamente em todos os *tweets* as celebridades retweetam o usuário.

Uma vez apresentados os usuários mais influentes dentro de cada medida, os autores investigaram como estas medidas estão correlacionadas. Para isso, a correlação relativa de todos os 6 milhões de usuários foram comparadas, considerando o valor de 0,5 uma alta correlação. Porém, junto com a alta correlação dos usuários surge o conceito de *"tied ranks"* (usuários que possuem o mesmo valor de correlação) entre usuários menos influentes, pois alguns deles não possuíam nenhum *retweet* ou menção. Para evitar este viés (*"tied ranks"*) foram considerados apenas os usuários mais influentes. Para isso, utilizou-se dois conjuntos: um com 1% e outro com 10% dos usuários mais bem classificados baseados na medida *in-degree*. Com esta limitação, os usuários mais bem classificados possuíam uma forte correlação nas medidas de influência *retweet* e menções, ou seja, os usuários que são frequentemente mencionados também são "retweetados' com frequência. Entretanto, a medida *in-degree* não está correlacionada fortemente com as outras medidas, visto que usuários considerados bem conectados não são necessariamente os mais influentes.

Mei et al. [33] identificaram as principais características para medir influência do usuário no Twitter, empregando o conceito de entropia e análise de correlação entre as características encontradas e os serviços de classificação de influência popular. Assim, os autores definiram onze (11) características que consideraram

serem importantes no processo de identificação de usuários influentes.

As características encontradas pelos autores são relacionadas a conta do usuário na rede social. Por exemplo, divisão entre ações de usuários (*retweets* e menções), cuja permite determinar usuários influentes de acordo com a quantidade de suas ações dentro da rede social, sendo considerado influente aquele que obteve mais interações publicando uma quantidade menor de *tweets*. Outras características presentes no trabalho são:

- **Conta verificada**, se um usuário possui uma conta verificada ou não. Este atributo é utilizado para identificar usuários famosos em diferentes domínios.
- **Idade da conta**, número de meses desde quando a conta foi criada. Os autores explicam que esta é importante, pois quanto mais antiga a conta mais provável é ela possuir um amplo alcance social.
- **Novos seguidores**, novos seguidores conquistados durante um período de tempo.
- **Novas menções**, a quantidade de menções ou respostas a um usuário durante um período de tempo.
- **Novos *retweets***, quantidade de novos *retweets* que um usuário obteve durante um período de tempo.
- **Número de listas públicas**, quantidade de listas públicas em que um usuário é membro.
- **Número de *retweets***, quantidade de *retweets* que um usuário possui atualmente.
- **Número de *tweets***, quantidade *tweets* publicados pelo usuário.
- **Número de seguidores**, total de seguidores que um usuário possui atualmente.
- **Razão entre seguidores/seguindo**, é a divisão da quantidade total de seguidores e seguindo do usuário. Se o valor resultante estiver mais próximo de 1 significa que este usuário segue de volta seus seguidores, caso mais próximo de 0 os usuários são considerados *spammers* ou *bots*.

A partir disso foi realizada a coleta de dados baseada em 100 usuários do Twitter e seus respectivos *tweets*. Após a coleta de informações destes usuários, os autores aplicaram o coeficiente de correlação de Pearson nos usuários com alta

correlação em suas características e aplicaram também o método de entropia para calcular os pesos para cada características escolhida.

Segundo os autores, os resultados mostraram que utilizando a correlação e o método de entropia foi possível obter uma eficácia boa para realizar predição de influenciadores. Dentro das características utilizadas, verificaram que o "número de listas públicas" é muito eficaz para demonstrar as pontuações de influência nos serviços de influência popular como o *Klout Score*. Além disso, foram analisadas outras características que podem ser coletadas através da API do Twitter, nas quais foram denominadas de características ocultas, e pode-se observar que popularidade, engajamento e autoridade são atributos sociais mais importantes para influência o usuário no Twitter.

Com base em trabalhos existentes, Zamparas et al. [52] propuseram diferentes métricas de influência que levam em consideração a análise do Twitter, bem como o comportamento dos seguidores de cada usuário. Segundo os autores, as métricas propostas refletem da melhor forma a influência e o comportamento de um usuário no Twitter. As métricas propostas foram as seguintes:

- **Afiliação**, lida com propriedades qualitativas de cada usuário identificado no Twitter. Por exemplo, relação entre o número de seguidores e o número de amigos.
- **Grau de interesse**, interesse dos seguidores pelas publicações do usuário.
- **Métrica de influência**, utiliza razão entre a afiliação do usuário e grau de interesse como cálculo de influência.
- **Média de influência de seguidores do usuário**, utiliza a média da Métrica de Influência para todos os seguidores de um usuário específico.

Para a realização dos experimentos os autores utilizaram páginas relacionadas a Cinema e Música, envolvendo um conjunto de 100 usuários.

A partir dos resultados, os autores observaram que para a métrica de afiliação, os usuários que seguiam a página relacionada à Cinema possuíam o valor de afiliação maior do que os usuários que eram seguidos pela página. Isso significa que usuários influenciadores tendem a possuir um menor valor de afiliação em relação a usuários não influenciadores, pois usuários considerados influenciadores não possuem o costume de seguir usuários que os seguem.

A segunda métrica, grau de interesse, resultou em um comportamento parecido com a métrica de afiliação. Mais precisamente, 5% dos usuários que eram seguidos pela página possuíam um alto valor para a métrica, isso porque existem diversos usuários que recebem muito interesse de seus seguidores através de *retweets*, menções e favoritos.

A métrica de influência possuiu um valor maior de 20% para os usuários que são seguidos pela página, sendo essa a maior diferença entre esta métrica em comparação com a grau de interesse. Isso ocorre porque a segunda leva em consideração a característica de seguidores e isso possibilita uma melhor identificação de usuários influentes, pois utiliza não só o conceito de interesse pelo conteúdo, mas também a sociabilidade do usuário.

Os resultados obtidos na última métrica, média de influência, mostraram que os valores para seguidores e usuários que seguem a página estão próximos. No geral, ficou concluído que os valores médios não podem diferir entre usuários diferentes, pois a média da métrica de influência se comportou de maneira parecida à métrica de influência do usuário. De acordo com os autores, quando o comportamento dos seguidores do usuário estão em estudo, os resultados são mais precisos, logo as conclusões são mais significativas e confiáveis.

Sun et al. [44] propuseram um modelo baseado em características do usuário como entrada para um modelo que envolve rede Bayesiana e o algoritmo *PageRank*, para identificar usuários influentes em redes sociais. Os dados coletados são pertencentes a rede social chinesa *Sina Weibo*. Dentre as características utilizadas se destacam os *retweets*, *tweets* e seguidores. Após a coleta dos dados, o modelo seleciona as características adequadas e constrói a rede Bayesiana para obter a possibilidade de um usuário ser influente. Por último calcula-se a influência do usuário adjacente com base no princípio do algoritmo *Page Rank*.

Analisando todas as características em conjunto foi encontrado uma grande relação entre a contagem dos *retweets* e contagem dos comentários. No entanto, não existia uma relação forte entre a autenticação e a contagem dos *tweets* para outras características. Os resultados atenderam ao objetivo proposto pelos autores de que a influência do usuário é determinada pela análise abrangente de múltiplas características e que a avaliação de influência deve ser analisada somente em determinados eventos ou tópicos para então ser significativa a avaliação de usuários influentes.

Lahuerta-Otero and Cordero-Gutiérrez [27] investigaram usuários influentes no Twitter para descobrir características de seus *tweets* através de uma ferramenta de pesquisa que combina teoria dos grafos e teoria de influência social. O estudo baseou-se na análise de hipóteses que são levantadas a partir de características de usuários e dos conteúdos publicados que são encontrados nas redes sociais. Os autores propuseram medidas nas quais acreditavam ter um grande impacto na influência de um usuário, tais como:

- **Influência**, que é um valor que varia de 1 a 100 de acordo com a popularidade do usuário medida nas suas redes sociais existentes;
- **Média de caracteres**, que é o número de símbolos (média) de um *tweet* do usuário;

- **Diversidade léxica**, o número de palavras únicas em um *tweet* do usuário dividido pelo total de palavras escrita no *tweet*;
- **Polaridade de sentimento**, que representa o sentimento de um *tweet* do usuário, podendo ser negativo, positivo ou neutro,
- **Seguindo**, quantidade de usuários que um indivíduo segue no Twitter.

Além destas medidas, também foram analizadas outras três (3) características encontradas em um *tweet*: *hashtags*, links e menções

A metodologia utilizada consistiu em coletar *tweets* dos usuários relacionados ao tema automobilístico, resultando em mais 3.853 usuários e de 30.000 *tweets*. Os resultados mostram que usuários influentes utilizam o artifício de *hashtags* e também menções a outros usuários em seus *tweets*. O número de palavras em seus *tweets* é menor do que aqueles com menos influência em uma mesma comunidade que um usuário influente. Além disso, outro aspecto encontrado foi a pouca utilização de links em suas publicações e em média possuem um grande número de amigos. Por fim, os seus *tweets* conseguem claramente expressar suas opiniões e sentimentos.

Díaz-Beristain et al. [17] analisaram quais fatores levam um usuário a obter mais seguidores dentro do Twitter, tornando-o influente na rede social. Para os autores, o significado de influência está relacionado com a capacidade de um indivíduo mudar a opinião ou percepção de outro. Para analisar a influência dos usuários do Twitter, os esforços se concentraram em investigar um padrão de influência e como os usuários considerados especialistas em um determinado campo podem promover o crescimento do número de seguidores de outros usuários utilizando a informação de *retweets*.

Tomando como base o trabalho de Lee et al. [28], os autores assumiram dezoito (18) tipos de categorias de interesse de usuários em rede social, definindo seis (06) classes de usuários de acordo com o número de seus seguidores. A Tabela 3.2 ilustra essas seis classes.

Classes	Intervalo de seguidores
Desconhecido	0 - 1,000
Comum	1,001 - 10,000
Bem Conhecido 1	10,001 - 100,000
Bem Conhecido 2	100,001 - 1,000,000
Bem Conhecido 3	1,000,001 - 10,000,000
Famoso	10,000,001

Para comprovar suas ideias, um experimento foi realizado com um usuário (*root*), onde extraíram informações sobre suas atividades (*tweets*, *retweets*, men-

ções e novos seguidores). O usuário escolhido possuía um total de 3.253 seguidores. O perfil deste usuário, desde sua criação, apresenta um crescimento de seguidores estável em torno de dois ou três novos por semana. Os *tweets* do usuário foram classificados como pertencentes a categoria de tecnologia, mas para o experimento seu comportamento foi modificado, através da diversificação de suas principais publicações de interesse, adicionando também imagens, vídeos e URLs de duas categorias: esporte e música.

O resultado do experimento mostrou que houve um comportamento regular do usuário *root* durante as semanas anteriores até a primeira fase do experimento que ocorreu durante quatro semanas, Logo na primeira semana, o usuário aumentou a quantidade de *tweets* e a diversidade de conteúdo, aumentando o número de seguidores. A análise também mostrou que ao utilizar algum tipo de mídia nos *tweets*, o usuário *root* chamou atenção de outros usuários e isso impactava diretamente no número *retweets* recebidos, como afirma [53]. Finalizando, os autores concluíram que não é fácil determinar fatores significantes que permitem um usuário do Twitter ganhar mais seguidores, pois tendem a ser mais estáveis. Contudo, algumas maneiras podem fazer com que um usuário "comum" ou "desconhecido" possa ganhar novos seguidores a partir do conteúdo e variedade de assuntos publicados.

Farahani et al. [21] investigaram comportamentos comuns dos usuários influentes através da aplicação de métricas, introduzidas por [37], como número de *retweets*, número de menções a outros usuários, número de *tweets* únicos "retweetado" por outros usuários, entre outros. Essas métricas são usadas para examinar a influência do usuário em diferentes dimensões para depois serem analisadas em suas relações. As métricas utilizadas pelos autores levam em consideração atributos que são extraídos diretamente dos usuários Twitter. A base de dados utilizada possuía informações de usuários, *tweets*, *retweets*, URLs e *hashtags* [29].

A análise de usuários influentes executou-se baseado nos 20 melhores usuários bem classificados após o cálculo das métricas. Após encontrar os usuários influentes o objetivo foi descobrir o padrão de comportamento e investigar a eficácia das diferentes métricas estudadas. Após a comparação das métricas foi aplicado um modelo de regressão para antecipar o comportamento dos usuários. Por fim os resultados mostraram que a métrica relacionada *tweets* originais é uma das mais importantes em termos de eficácia do usuário, tendo um impacto direto na influência de um usuário do Twitter.

Asadi and Agah [6] analisaram influência de forma quantitativa a partir de dados do Twitter, baseando usuários em dois fatores: aqueles que eles seguem e como eles são ativos. Os autores partem da hipótese de que para um usuário ser um influenciador ele deve tentar engajar seu público, porém não há uma definição clara de como isso deve ser alcançado. Para tentar descobrir qual estratégia o

influenciador utiliza para ganhar audiência, foram propostas quatro medidas: número de seguidores, usuários favoritos, nível de atividade e razão entre número de seguindo e seguidores. A partir das medidas calculadas, os usuários foram divididos em três categorias: famosos, comuns e desconhecidos.

Dentre as análise realizadas podem se destacar que o nível de atividade do usuário possuía pouco efeito para quantificar a influência de um usuário, a porcentagem de *tweets* que eram *retweets* ou respostas parecia ter um efeito muito maior sobre o tamanho da audiência. Outro fator que os autores destacam é o fato de influenciadores gastarem um quantidade de tempo "retweetando" e respondendo outros usuários de alta influência para assim obter um grau de influência maior.

3.3 Discussões

Um dos maiores desafios encontrados na literatura é como definir e mensurar influência em redes sociais. Neste Capítulo foram apresentados trabalhos que encontraram diferentes formas de definir (utilizando contexto diferentes ou baseando-se em trabalhos anteriores) e mensurar influência. A ideia geral foi utilizar métricas nas quais tentam representar fielmente os usuários e seus comportamentos dentro da rede social para, assim, poder identificar influência e determinar como ocorre a difusão de informação dentro da rede social.

A Tabela 3.1 sumariza os trabalhos relacionados apresentados neste Capítulo.

Tabela 3.1: Trabalhos relacionados

Autores	Formas de identificação		Métricas Utilizados						Características Utilizadas				
	Conteúdo	Topologia	*Degree*	*Closeness*	*Betweenness*	*Eigenvector*	*Page Rank*	Outras métricas	*Hashtags*	*Retweets*	*Tweets*	Seguidores/Seguindo	Menções
Dubois and Gaffney [18]		✓	✓			✓		✓					
Räbiger and Spiliopoulou [39]		✓	✓	✓	✓	✓							
Al-Garadi et al. [2]		✓	✓				✓	✓					
Al-garadi et al. [1]		✓						✓					

Zhuang et al. [58]		✓					✓						
Alp and Öğüdücü [5]		✓					✓						
Yang et al. [50]		✓		✓	✓								
Cha et al. [13]	✓									✓		✓	✓
Mei et al. [33]	✓								✓	✓	✓	✓	✓
Zamparas et al. [52]	✓									✓	✓		✓
Sun et al. [44]	✓									✓	✓	✓	
Lahuerta-Otero and Cordero-Gutiérrez [27]	✓								✓		✓	✓	✓
Díaz-Beristain et al. [17]	✓									✓	✓	✓	✓
Farahani et al. [21]	✓								✓	✓	✓		
Asadi and Agah [6]	✓									✓	✓	✓	

Percebe-se na Tabela que os trabalhos baseados no conteúdo do usuário empregam diferentes combinações de características, a fim de aumentar a quantidade de identificações de usuários influentes. Os autores que utilizam esta forma de identificação afirmam que, para obter eficiência em seus métodos, é necessário empregar diversos tipos de características que o Twitter fornece. A explicação pode ser embasada pelo fato que as características que envolvem usuários e conteúdos são facilmente modificados e são muito voláteis quando observando comportamento a longo prazo. Ou seja, um usuário considerado influenciador pode não receber muitos *retweets* quanto um usuário comum se for avaliado apenas um tópico. Mas este tipo de característica se torna importante quando analisada em um conjunto grande e diversificado de tópicos, pois assim podem ser encontrados padrões que diferenciam usuários comuns e influentes.

Dentro desta forma de identificação. a característica mais utilizada é o *retweet*. Segundo os autores, esta é a que melhor representa o comportamento de usuário na rede social, tanto em influência quanto em difusão de informação dentro da rede social, pois a ação de retweetar é compartilhada em forma global com toda rede. Entretanto, existem características que não possuem uma grande eficiência ao serem avaliadas sozinhas, como, por exemplo as *hashtags*. Todos estudos não as utilizam sozinhas, pois não representam verdadeiramente comportamentos em rede social. As usam de forma mesclada com outras características do mesmo conjunto de identificação.

Já os trabalhos que utilizam a topologia de rede para realizar a identificação de usuários influentes possuem uma vantagem principal em relação à análise

de conteúdo, que é o fato de analisar de forma global usuários e seus posicionamentos e variações dentro da rede ou de uma comunidade. Esta abordagem também facilita a identificação dos usuário chamados *bots*. O ponto negativo desta abordagem é o poder de processamento necessário pelas métricas quando se avaliando redes consideradas grandes. De forma geral, as métricas se tornam custosas quando são analisadas redes que representam o mundo real. Outro aspecto limitante é ter que avaliar a rede em diferentes pontos de vista, utilizando para isso diversas métricas, visto que ao utilizar somente as mais básicas a rede pode não representar verdadeiramente o mundo real.

Ao contrário do que acontece em formas de identificação que utilizam o conteúdo, analisar métricas que pertencentes a topologia do usuário possuem mais diversidade para determinar usuários influentes e difusão de informação. Conforme mostrado na Tabela, alguns autores mesclam tipos de métricas existentes. Uma explicação para não existirem muitos trabalhos que utilizem a união de várias métrica é pelo motivo que elas conseguem retratar mais verdadeiramente uma rede social. Para isso, é comumente utilizado métricas que envolvem a análise dos "vizinhos" do usuário, por exemplo, *Degree centrality* e *Page Rank*.

Outro ponto a ser notado na Tabela é o campo "Outras Métricas". Nele aparecem três trabalhos que fizeram uso de outros tipos de recursos, ao invés de utilizar somente métricas simples ou análises consideradas "rasas", como, por exemplo, avaliar somente uma camada de rede. Dentre esses recursos utilizados, que melhoram a identificação ou visualização da rede usuários influentes, podem ser citados: a separação em multicamadas [2], a detecção de comunidades [39] e o método baseado no *k-core* que é ponderado com base nas interações entre os usuários [1].

Embora não visível na Tabela, percebe-se que nos trabalhos apresentados existem lacunas a serem estudadas. O problema não é em relação as métricas empregadas, mas sim no fato de que para determinados eventos ou temas o conceito de influência pode variar e que nem sempre um usuário comum pode ser influente no tema. Outro aspecto faltante é uma análise do conceito de influência em redes que representam o mundo real, visto que quando há uma grande quantidade nós e arestas na rede, o poder de processamento aumenta. Assim, as soluções apresentadas, de forma geral, trabalham com redes menores ou empregam métodos que processam somente as chamadas "small worlds".

Esta dissertação compartilha da visão que é possível mensurar a influência e a difusão de informação mesclando características topológicas e baseadas no conteúdo, pois isso permite avaliar contextos diferentes e empregar características que podem incrementar outras. Nesta dissertação, como uma das propostas, será realizado um estudo utilizando estes dois tipos de características em conjunto e avaliando-as para saber se há uma correlação entre si. Também espera-se de-

monstrar que métricas muitas vezes utilizadas para identificar usuários influentes podem não refletir verdadeiramente em qualquer tipo de conjunto de dados.

Capítulo 4

Metodologia e Experimentação

Este Capítulo descreve os aspectos, passos e ferramentas essenciais para alcançar o objetivo proposto de investigar usuários e sua influência dentro das redes sociais. Para isso serão descritos a metodologia em modo geral, as características utilizadas (usuário e topologia), ambiente utilizado para coleção das bases de dados e descrição geral das particularidades de todas as bases utilizadas.

4.1 Metodologia e Implementação

Para analisar e demonstrar que as atuais métricas para identificação de usuários influentes não são aplicáveis em qualquer cenário ou evento em redes sociais, uma metodologia de avaliação foi criada, englobando passos que vão da coleta de um evento (que no caso desta dissertação são as *hashtags* que se encontram nos tópicos de tendência) até à análise da base de dados que respondam as questões de pesquisa introduzidas no Capítulo 1.

Vale lembrar que o foco desta dissertação é a rede social Twitter e suas interações sociais (curtidas, *retweets*, menções e *tweets*), que permitem inferir o comportamento de cada usuário dentro da rede. No caso desta dissertação, serão utilizadas para identificar usuários influentes e analisar o poder difusão de informação dentro da rede social, que cada usuário possui, baseados nos eventos coletados, os chamados tópicos de tendências (*trend topics*). Para analisar esses comportamentos são utilizadas métricas e/ou características individuais e coletivas.

A Figura 4.1 ilustra a visão geral da metodologia.

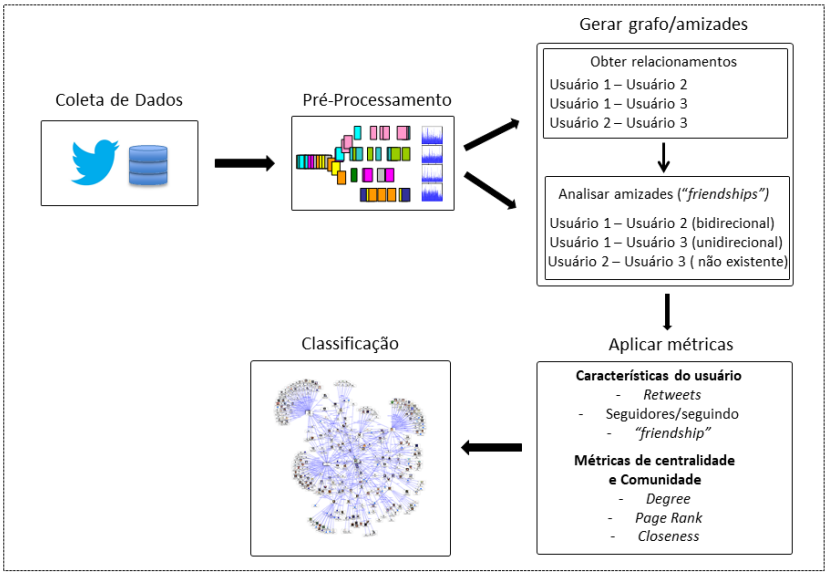

Figura 4.1: Visão geral da Metodologia proposta

4.1.1 Etapas da Metodologia

Coleta de dados

A primeira etapa da metodologia é a coleta de dados. Basicamente, são coletadas *hashtags* que possuem alto índice de menções e, logo, fazem parte dos *trend topics* do Twitter. Para tanto, a coleta emprega métodos que são utilizados através de APIs que são disponibilizadas pelo Twitter algumas delas apresentada na Tabela 4.1. Nesta dissertação é utilizada a *Search API*, na qual faz buscas de histórico de *tweets* dos usuários.

Os métodos listados na Tabela 4.1 são utilizados na construção desta metodologia envolvendo tópicos de tendência, informações sobre usuários e as publicações realizadas, porém existe uma variedade de outros métodos que são utilizados para diversas finalidades. Elas utilizam o protocolo HTTP para fornecer aos seus clientes determinadas funcionalidades descritas por conjuntos de recursos que podem ser acessados remotamente por meio de requisições HTTP comuns. A partir disso, são recebidas coleções de dados estruturados no formato JSON (*JavaScript Objection Notation*) que possuem o conteúdo em formato texto codificado com algum esquema de caracteres. Esta API fornece acesso a vários métodos e recursos, porém muitos deles disponíveis necessitam de autenticação para que sejam utilizados.

Tabela 4.1: Lista de métodos da *Search API*

Método	**Objetivo**
API.user_timeline	Retorna os 20 recentes status publicados pelo usuário
API.get_user	Retorna informação sobre o usuário
API.show_friendship	Retorna de forma detalhada as informações entre dois usuários
API.followers	Retorna os seguidores de um usuário de forma ordenada
API.trends_available	Retorna os tópicos de tendência do local que é passado para a API

De modo prático, foi desenvolvido um *crawler* que utiliza a API padrão do Twitter para fazer autenticação e assim coletar os dados. O *crawler* foi desenvolvido na linguagem de programação Python, versão 3.5, através da biblioteca *tweepy*. Nesta dissertação, os dados coletados foram relacionados a dois (02) tópicos de tendências nos quais envolvem política e racismo, temas considerados polêmicos. A coleta destes dados ocorreram em dias e meses distintos, mas sempre com grande participação de usuários, através das *hashtags* que estavam entre os principais *trend topics* naquele momento.

Pré-Processamento

Após a fase de coleta de dados é realizado o pré processamento dos conjuntos de dados que foram coletados. Neste passo são organizados e retirados dados que não servem para uma futura construção da rede e/ou análise dos usuários. Em geral, são eliminados dados de usuários que não possuem interações, ou seja, não receberam e também não atribuíram nenhum *retweet*.

Este procedimento permite uma melhor visualização da rede, pois apenas existirão usuários que possuem interações e assim poderão ser analisadas a influência e a difusão de informação em conjunto.

Obtenção de Relacionamentos

Nesta etapa são criados os relacionamentos dos usuários dentro dos dados coletados. Estes relacionamentos são obtidos através da interação entre os usuários e *retweet*, pois são determinantes para analisar a influência de um certo usuário e também a difusão de informação dentro da rede. A extração desta relação é obtida a partir dos usuários que são pré-processados, onde para cada usuário existe a informação se o mesmo "retweetou" uma publicação e quem é o autor original da mensagem, incluíndo todas as credenciais do usuário. No Twitter, os

usuários podem ser representados pelo nome de sua conta ("*screen name*") ou pelo seu identificador (ID). Nesta dissertação, é utilizado o ID como forma de identificação do usuário, um atributo não volátil que preserva a identidade dos usuários.

Neste etapa também são aplicados, em paralelo, outros dois passos. O primeiro é a geração completa do grafo das interações dos usuários e o segundo é a verificação de amizades entre os usuários que possuem interações. Quando um usuário "retweeta" uma publicação de outro usuário é verificado o conceito de amizade (*friendship*), ou seja, utilizando-se a API do Twitter consegue-se obter o tipo de relação que os dois usuários possuem dentro da rede social. Nesta dissertação serão utilizados os seguintes casos: seguir, seguido, mútuo e não possuem amizade.

- Seguir: usuário que retweetou segue o usuário que publicou *tweet*.
- Seguido: usuário que retweetou é seguido pelo usuário que publicou o *tweet*
- Mútuo: há reciprocidade na relação dos dois usuários.
- Não possuem amizade: nenhum dos dois usuário segue ou é seguido.

A partir disso, é gerado um grafo de relacionamentos dentro daquele determinado tópico. Neste caso, para esta visualização destes grafos, é utilizado o software Gephi[1], para posteriormente utilizar as métricas propostas.

Aplicação das Métricas

Após a construção da rede, a etapa de aplicação de métricas é realizada para a identificação inicial dos usuários mais importantes dentro do grafo. Para descobrir quem são os principais usuários da rede é aplicada a métrica *Degree Centrality*, que retorna a quantidade de interações que o usuário obteve somente na coleta dos dados (*In-Degree*) e a quantidade de vezes que o usuário interagiu com outros (*Out-Degree*).

Outras métricas que envolvem topologia e características individuais de cada usuário também são aplicadas, sendo esta última métrica podendo ser aplicada através de diversificações com outras características, por exemplo, *retweets* mais seguidores e seguindo. Por último, é proposta uma métrica que envolve os dois tipos de identificação utilizados mais a análise realizada com amizades dos usuários (*friendship*).

As primeiras métricas aplicadas foram relacionadas a topologia do usuário, pois pretende-se analisar de um aspecto geral como estão organizados os usuário

[1] https://gephi.org/

em forma de rede social. Para isso são utilizadas métricas que envolvem informações sobre vizinhos, ou seja, quantidade de nós próximo de um usuário específico ou o quanto importante é um usuário dentro da rede.

Entretanto, são analisadas em conjunto métricas que relacionam caminho mínimo, ou seja, saber o quanto distante está um usuário que espalha informações dentro da rede em relação ao usuário influente. Este tipo de métrica ajuda a determinar como se comportou um usuário dentro de um evento coletado.

As métricas foram aplicadas na ordem em que são listadas abaixo.

Topologia

- *In-degree*: número de arestas direcionadas ao nó, neste caso o número de *retweets* que o usuário possuiu durante aquele evento coletado.
- *Out-degree*: número de arestas que o nó encaminha aos outros, ou seja, quantidade de *retweets* que um usuário concedeu a outros.
- *Page Rank*: assume que a contribuição de centralidade dos vértices não é a mesma e deve ser penalizada proporcionalmente de acordo com a quantidade de vizinhos
- *Closeness*: métrica baseada em caminho que calcula a distância de um nó para todos os vértices da rede.
- *Betweenness*: métrica também baseada em caminho, em que o cálculo leva em consideração os caminhos mínimos, onde para cada par de vértices em um grafo conectado, existe pelo menos um caminho mais curto entre os vértices

Após utilizadas as métricas de topologia são aplicadas métricas que utilizam características do usuário, onde serão analisados usuários que possuíram uma grande importância através de características topológicas e avaliados se este comportamento predomina em escalas menores, por exemplo, avaliar somente o usuário em questão.

Assim como visto anteriormente, as métricas neste ponto possuem a finalidade de avaliar diferentes perspectivas de um usuário. Neste caso, as métricas são avaliadas nos seguintes quesitos: popularidade, influência e difusão de informação. Popularidade analisa a quantidade de seguidores de um usuário, não levando em conta sua influência sobre eles. Influência analisa como a popularidade aspecto intervém sobre os seguidores. A Difusão de informação analisa o poder de um usuário de espalhar informações próprias ou de terceiros dentro da rede social.

As métricas foram aplicadas na ordem em que são listadas abaixo.

Características do Usuário

- *Follower Rank* [35]: métrica de popularidade que é baseada na quantidade de seguidores e seguindo de um certo usuário, para o cálculo é utilizado uma divisão entre os valores. Podendo ser representada na equação 4.1 a seguir.

$$FollowerRank(i) = \frac{Followers}{Followers + Followees} \tag{4.1}$$

- Popularidade [4]: métrica que é calculada baseada no número de arestas que um certo usuário obtém, por exemplo, quantidade de *retweets* que recebe. Como exemplifica a equação 4.2

$$Popularity(a) = 1 - e^{\lambda F(a)} \tag{4.2}$$

- *Retweet Impact* [37]: esta métrica estima o impacto dos *tweets* dos usuários, usando como referência os *retweets* obtidos durante a publicação. Como mostra a equação 4.3

$$RI(i) = RT1. \log(RT2) \tag{4.3}$$

  Onde *RT1* significa a quantidade de *retweets* recebidos por outros usuários em uma publicação específica, enquanto *RT2* significa a quantidade geral de *retweets* recebidos de outros usuários para todas as publicações realizadas por um autor.

- Difusão de Informação [37]: métrica que estima a possibilidade dos *tweets* de usuários influenciarem entre os seus seguidores, a equação 4.4 demonstra a métrica.

$$DI(i) = \log(F1 + 1) - \log(F2 + 1) \tag{4.4}$$

  As variáveis *F1* e *F2* significam os *tweets* que foram publicados utilizando o mesmo tema ou assunto após a publicação inicial de um usuário, através de seus seguidores e seguindo, respectivamente.

As métricas mostradas possuem finalidades diferentes entre si. Por exemplo, as de topologia baseiam-se em usuários vizinhos (*In-degree*, *Out-degree*, e *Page Rank*), caminhos entre usuários (*Closeness* e *Betweenness*) e o conceito de detecção de comunidades. Quanto as características do usuário, encontram-se as que utilizam os seguidores como base (*Follower Rank*, e Popularidade), *retweets* (*Retweet Impact*) e seguidores mais *retweets* (Difusão de Informação).

A última métrica pro, noposta, se preocupa em investigar apenas, no nível de interação, se a importância destes usuários se deve aos seus seguidores. Com isso

foi realizada uma verificação de existência de amizades entre as contas coletadas durante o evento. O *script* desenvolvido tem como parâmetros o usuário que realizou uma publicação sendo este chamado de *source* e os usuário que retweetaram a publicação *target*. Para cada métrica é gerado um grafo de relacionamento onde destaca os usuários mais importante para métrica em questão.

Classificação

Como resultado da aplicação das métricas, os usuários mais importantes são destacados nos dados coletadas, podendo serem considerados usuários influentes e/ou difusores de informação dentro da rede. Para isso, os usuários são classificados em rankings dos mais importantes para os menos importantes, dentro cada tópico coletado, analisando sempre se há interferência das métricas em características diferentes. Ao final, são destacados os usuários que possuem presença em todos os tópicos coletados, respondendo as questões de pesquisa relacionadas.

4.2 Bases de Dados

Conforme detalhado na metodologia, nesta dissertação foram coletadas duas bases de dados (Tabela 4.2) durante diferentes períodos, pertencentes a temas polêmicos como: política e racismo. A escolha desses temas se deve aos mesmos terem ficado em alta nos tópicos de tendência durante dias e semanas, resultando em diversos usuários usando *hashtags* que simbolizavam o evento em questão. Foram coletados *tweets* e consequentemente *retweets* atribuídos a essas publicações. Outros dados coletados foram: nome de usuário, ID, data de criação do *tweet*, curtidas, presença de algum tipo de mídia dentro do *tweet* (imagens, vídeos ou links).

As bases de dados são detalhadas nas subseções a seguir e apresentadas em ordem cronológica.

Tabela 4.2: Tabela de Bases Coletadas

Base	**Usuários (vértices)**	**Interações (arestas)**
Racismo	50,045	62,065
Política	44,258	104,469

4.2.1 Racismo

A base de dados relacionada à racismo foi coletada no período entre os dias 26 e 27 de Novembro de 2017. Trata-se do caso de racismo envolvendo a filha de

um casal famoso e tornou-se um dos assuntos mais comentados na época. Uma característica sobre esse evento é que o mesmo não fez necessariamente uso de *hashtags* para estar entre os *trend topics*, mas a *hashtag* e o nome dos famosos permaneceram em alta durante o período citado acima.

Para não envolver outros casos semelhantes ao que estava sendo coletado, o script desenvolvido utilizou combinações de *hashtags* e nomes, de forma que as ocorrências encontradas apenas trouxessem esta combinação. As strings de busca para este tópico foram: "Bruno Gagliasso *and* racismo" *or* "Titi *and* racismo". A coleta dos dados resultou em mais 66 mil ações de usuários entre *tweets* e *retweets*. Entretanto, como a finalidade é avaliar somente as interações para representar mais tarde grafos, reduziu-se para 62 mil com somente interações (*retwets*). O número de usuários (vértices) existentes alcançou 50.045 e o número de interações foi de 62.065 (arestas).

4.2.2 Política

A base de dados relacionada à política foi coletada no período de 24 à 28 de Janeiro de 2018, envolvendo o julgamento de um ex-presidente da república. Nestas datas, diversas *hashtags* foram utilizadas como forma de expressar apoio ou repúdio ao mesmo. Nesta dissertação, as *hashtags* escolhidas foram #EleicaoSemLulaEFraude *or* #CadeAProva e #MoluscoNaCadeia *or* #CondenaTRF4, ambas com o objetivo de representar um dos lados políticos.

Como a base foi coletada utilizando todas as quatro (04) *hashtags* em conjunto, a mesma possuía alta polaridade entre os usuários, chamados nesta dissertação de pró ou contra. Vale ressaltar que a base definida como "contra" foi maior na quantidade de usuários que participaram do tópico coletado. A principal explicação para este fato se deve a quantidade de dias em que as *hashtags* se mantiveram nos tópicos de tendência da rede social.

4.2.3 Ambiente de Experimentação

Os experimentos realizados para a elaboração desta dissertação foram executados utilizando duas máquinas de configurações de diferentes. A primeira sendo utilizada para coleta dos dados, pré processamento e criação inicial das interações dos grafos em paralelo com análise das amizades entre usuário, possuindo as seguintes configurações: Intel(R) Xeon(R) 2.53GHz, com 9GB de memória RAM, disco de rígido de 1TB, plataforma Linux, distribuição Debian. A segunda máquina utilizada nos experimentos para a visualização dos dados em grafos e análise das métricas propostas possui as seguintes configurações: Intel Core I5, com 8GB de memória RAM, disco rígido de 500GB e plataforma Linux, distribuição Mint.

4.3 Discussões

Conforme apresentado, este Capítulo possui como objetivo detalhar e discutir os dados, a metodologia e as fases que possuem para que possa ser realizada a análise sobre os dados obtidos durante os experimentos. Além de detalhar também o processamento utilizado para se chegar a um resultado final.

A metodologia utilizada possui resumidamente as fases de coletas dos dados, pré processamento e análise dos dados gerados. Entretanto, dentro de alguma dessas fases existem tarefas que são executadas paralelamente, como gerar os grafos de interações dos usuários coletados. Dentro deste passo é executado também uma tarefa na qual é denominada como análise de amizades (*friendship*), quando o grafo está sendo gerado esta análise ocorre paralelamente, pois será utilizado posteriormente como parte das métricas propostas nos dois contextos apresentados na seção 4.1.1. Um dos motivos de realizar esta tarefa em paralelo é pelo fato de que usuários constantemente modificam seus comportamento nas redes sociais, tais como: deixar de seguir um usuário, deletar sua conta, tornar sua conta privada o que torna impossível realizar qualquer tipo de análise na mesma.

Realizada esta fase de coleta de amizade existente entre os usuário dentro do evento, foram mostradas as interações em forma de grafos para todas as duas bases de dados geradas. A representação neste formato permite aplicar as métricas definidas e assim identificar usuários que podem ser considerados influentes ou difusores de informação. A métricas utilizadas nesta dissertação envolvem características diversas, onde podem se destacar em dois grupos, características do usuário e a representação em forma de topologia da rede social.

A principal finalidade de utilizar as duas representações será para avaliar em conjunto as duas formas existentes para identificar usuários influentes e difusores de informação dentro da literatura, porém as métricas podem não representar fielmente quando aplicadas em contextos do mundo real, neste caso utilizando redes sociais, como será apresentado no próximo capítulo.

Capítulo 5

Resultados

Neste capítulo são descritos os resultados encontrados nas bases de dados geradas a partir das análises baseadas nas métricas definidas no capítulo anterior. Os resultados serão detalhados individualmente em cada base de dados para ao final identificar a existência de alguma particularidade, seja nas métricas empregadas ou nos usuários analisados, as métricas que envolvem características do usuário são avaliadas em conjunto com as características topológicas. Com esta análise serão respondidas as questões de pesquisas formuladas no **Capítulo 1** como forma de alcançar o objetivo proposto nesta dissertação. E ao final de cada base de dados apresentada são discutidas suas particularidades e o que diferem tratando os conceitos de influência e difusão de informação.

5.1 Racismo

Como apresentado no capítulo anterior, esta base de dados foi gerada a partir da coleta de dados relacionadas a um tópico de tendência que obteve destaque na época, com isso usuários, *tweets* e formas de interações foram processadas. A base gerada possui mais de 50 mil usuários e mais 62 mil interações, neste caso *retweets* entre estes usuários. O assunto publicado e comentado por diversos é relacionado a racismo, considerado um tema polêmico, ou seja, possuindo diversas interações entre os usuários.

As interações entre os usuários foram avaliadas em forma de grafo no qual são organizados em forma de comunidades, ou seja, ao utilizar o algoritmo de *Louvain* [9] os usuários irão ser organizados dentro de comunidade, nas quais são identificadas pelas diferentes cores. Esta forma é utilizada nesta dissertação pois é a maneira encontrada sendo mais explicativa para analisar e mostrar as interações entre os usuários da base e também as diferenças entre as métricas propostas.

5.1.1 Visão geral

Conforme é apresentado na Figura 5.1, são identificadas comunidades que foram encontradas de acordo com as interações dos usuários durante o evento, onde diversos usuários podem estar presentes em uma ou várias comunidades simultaneamente, ou em alguns casos não estarem presentes em comunidades, porém nesta dissertação não foi considerado este último caso. Os nós mais destacados na imagem são usuários que possuíram mais importância dentro do evento coletado, ou seja, foram os que receberam mais *retweets*.

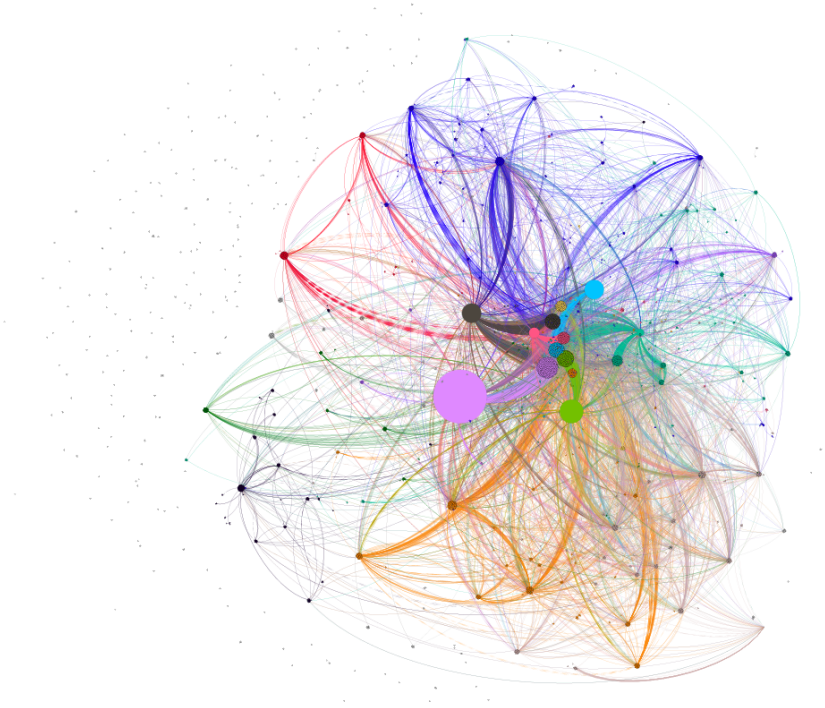

Figura 5.1: Visão geral da base 1 relacionando os usuários mais retweetados

Com isso, foram aplicadas as métricas propostas no capítulo anterior, primeiramente as que representam a topologia do usuário e posteriormente as que utilizam características provenientes do usuário.

In-degree

Esta métrica, como explicada anteriormente, define os usuário que possuíram mais importância dentro evento coletado, ou seja, são calculadas apenas as interações que recebe de outros usuários, nesta dissertação são utilizado os *retweets*

como forma de obter a influência e analisar também a capacidade de difundir informação que os usuários possuem, pois utiliza a quantidade de seguidores aplicada como características de usuário. Portanto, o usuário que obteve o maior destaque nesta métrica encontra-se na parte mais central da Figura 5.1, possuindo uma coloração mais roxa, conectado em outras comunidades vizinhas. Os usuários vizinhos que se conectam em sua comunidade também possuem um *in-degree* alto.

Os usuários representados no grafo possuem diferentes valores para a métrica calculada dentro de intervalo entre 1 e 18,935 interações recebidas. Conforme menor a quantidade de *retweets* recebidos maior fica quantidade de usuários repetidos com o mesmo retweets, comprovando o conceito de *tied ranks* como apresentado em [13]. Para esta métrica, os usuários com menores *in-degree* não são considerados nesta primeira análise. A Tabela 5.1 identifica os usuários com os maiores *in-degree* encontrados dentro da base.

Tabela 5.1: Classificação dos usuários baseados na métrica *in-degree*

	ID usuário	***In-degree***
1	usuário#8653	18,935
2	usuário#7514	8,182
3	usuário#3383	6,644
4	usuário#2479	6,633
5	usuário#7887	3,424
6	usuário#8802	2,203
7	usuário#7119	2,080
8	usuário#7748	1,277
9	usuário#5008	1,157
10	usuário#1505	995

Conforme apresentado na Tabela, o primeiro usuário possui o *in-degree* de 18,935 devido os *retweets* recebidos e que foram coletados utilizando a metodologia, possuindo uma grande disparidade em relação ao restantes dos usuários classificados. Uma explicação para este fato do "usuário#8653" ter sido um dos primeiros *tweets* relevantes publicados envolvendo o assunto. Associa-se a isso o fato de utilizar mídias (fotos, vídeos ou URLs), o que alavancaram ainda mais seus *tweets*, difundido-os pela rede social.

Dentro desta classificação encontram-se também diferentes tipos de perfis, como políticos e noticiários. A partir do usuário 11 (onze) em diante a quantidade de *in-degree* de cada usuário diminui consideravelmente, o que mostra que as grandes interações e divulgações de conteúdo sobre o evento coletado permaneceram praticamente com os 5 (cinco) primeiros usuários da classificação obtida.

Outro fator analisado são as diferentes comunidades que estes primeiros usuário se concentram. Praticamente cada um conseguiu mover uma massa de usuários e criar uma comunidade de acordo com suas publicações, como apresenta a Figura 5.1

É importante ressaltar também que esta quantidade de *in-degree* são apenas os *retweets* que o script desenvolvido conseguiu coletar, utilizando os tópicos de tendência mostrados no capítulo anterior. Portanto, existem casos em que o número de *retweets* recebidos pode ser maior que o *in-degree* calculado, pois as publicações permanecem disponíveis para todo o público.

Para entender mais claramente a métrica é necessário mostrar outras características que ajudem a esclarecer o comportamento do usuário dentro da rede social. Nem todos os usuários considerados influentes mostrados na tabela são exclusivamente famosos ou possuem muitos seguidores. Essas características são discutidas ao final do capítulo.

Out-degree

A métrica *out-degree* identifica os usuários que mais se relacionaram com um ou mais usuários dentro da rede geral, ou seja, um primeiro usuário que concedeu diversos *retweets*, sendo o responsável por divulgar *hastags* ou *tweets* que representaram aquele evento específico. Nesta base de dados, os usuários possuíram uma quantidade baixa de *out-degree*, se comparado diretamente com os valores de *in-degree*. Uma explicação para o fato se deve à quantidade de vezes em que se permite retweetar uma publicação original de um certo usuário. Porém, ainda assim é permitido retweetar outras diferentes publicações.

Como é apresentado na Figura 5.2, poucos usuários possuíram um valor alto para *out-degree*. Apenas um usuário é o mais destacado dentro da rede, localizado na parte inferior direita do grafo possuindo uma coloração bege. Este usuário "usuário#9337" não possui nenhum *retweet*, portanto tem o valor na métrica *in-degree* zerada, pois não criou nenhuma publicação no período do evento coletado, apenas *retweets* a outros usuários.

Um dos pontos importantes analisado foi a existência de interações com outros usuários considerados importantes dentro da rede (*in-degree*), pois o usuário "usuário#9337" possui conexão com os principais apesar de não estar localizado dentro da mesma comunidade que os demais.

Outro fator é a existência de usuários dentro da mesma comunidade (bege) que possuem os principais valores de *out-degree*. Nos dez (10) primeiros (Tabela 5.2), existem seis (6) que são pertencentes a mesma comunidade. Analisando os primeiros classificados, são encontrados sete usuários pertencentes a mesma comunidade, significando que ela é uma das responsáveis diretas em difundir dentro da rede o conteúdo gerado por outros usuário durante o evento coletado.

Figura 5.2: Visão geral da base 1 relacionando os usuários mais retweetados

Tabela 5.2: Classificação dos usuários baseados na métrica *out-degree*

	ID usuário	***Out-degree***
1	usuário#9337	75
2	usuário#3407	24
3	usuário#8674	23
4	usuário#3789	22
5	usuário#2853	22
6	usuário#1752	19
7	usuário#4867	18
8	usuário#1959	17
9	usuário#1947	17
10	usuário#2978	17

Porém, analisando individualmente, perecebe-se que estes mesmo usuários possuem poucos seguidores. Nenhum deles possui mais de cinco mil. Também são percebidos os perfis criados recentemente e ligados ao tópico coletado com o objetivo de interagir com o máximo de usuários, a fim de se tornar notável dentro da rede social.

Betweenness e Closeness

Por se tratarem de medidas que determinam caminho mínimo dentro da rede estudada, as duas métricas (*Betweenness e Closeness*) serão avaliadas em conjuntos neste tópico, pois apresentam resultados parecidos na classificação de ambos.

Usuários que possuem um alto valor na métrica de *in-degree* também estão entre os maiores na métrica de *betweenness*, cujo objetivo é encontrar usuários considerados mais centrais, ou seja, entre dois pontos (usuários) existe um certo usuário que serve como ponte entre os dois, calculando assim o caminho mínimo entre dois usuários. Usuários mais centrais funcionam como uma espécie de ponte para outros usuários conhecerem e, a partir disso, alavancarem outras publicações.

Em relação a métrica *closeness*, diversos usuários possuem valores iguais, significando que tem uma distância média próxima em relação a todos os outros usuários da rede. Por exemplo, o usuário "ddlloren", que obteve o maior valor na métrica *betweenness*, também recebeu o maior valor em *closeness*, pois este usuário e outros, que estão com os maiores valores na métrica, possuem um grande índice de interações, ou seja, muitos *retweets*.

Fica mais claro observar estes valores quando são comparados com a quantidade de *in-degree* que receberam a partir de todas as publicações realizadas somando com todos os *retweets* gerados por elas, pois nem todos os usuários mais bem classificados realizaram publicação somente uma vez, e a metodologia proposta classifica os usuários somando todos os seus *retweets* recebidos em decorrer das publicações realizadas. Este fator pode ser decisivo para que um usuário se torne mais central em relação a outros, caso sua influência for avaliada a partir de publicações únicas.

Page Rank

Esta métrica possui uma grande importância em identificar usuários influentes na rede social, pois não utiliza diretamente valores de centralidade totalmente, mas sim a utiliza penalizando proporcionalmente de acordo com o número de vizinhos do usuário.

Para esta métrica, os usuários mais bem classificados continuaram possuindo valores altos, alterando pouco a questão de influência entre si, exceto ao fato

de o usuário "usuário#2606" estar entre os mais bem classificados como mostra a Tabela 5.3. No entanto, o mesmo possui na métrica de *in-degree* o valor 1 (um), recebendo apenas uma interação na única publicação realizada durante todo o evento coletado. Porém, como dito anteriormente, esta métrica possui o objetivo de penalizar proporcionalmente a questão da centralidade ou influência entre usuários bem conectados para os menos conectados. Neste caso, o usuário "ddlloren", que obteve o *in-degree* de 8.182, possui uma interação, sendo esta com o usuário "usuário#2606" através de um *retweet*.

Tabela 5.3: Classificação dos usuários baseados na métrica *Page Rank*

	ID usuário	***In-degree***	***Page Rank***
1	usuário#8653	18,935	0,144
2	usuário#7514	8,182	0,055
3	usuário#2479	6,633	0,051
4	usuário#2606	1	0,047
5	usuário#3383	6,644	0,045
6	usuário#7887	3,424	0,023
7	usuário#8802	2,203	0,012
8	usuário#7119	2,080	0,012
9	usuário#5008	1,157	0,007
10	usuário#7748	1,277	0,006

Portanto, o usuário "usuário#2606" obteve uma grande importância ao ser retweetado por outro que também é considerado influente dentro, dividindo assim sua importância com os demais. Para exemplificar, seria o mesmo princípio de um usuário famoso interagir com outro menos conhecido ou não conhecido dentro da rede social. O último receberia uma importância igual, pois é considerado tão influente quanto. Porém, mais a frente será analisado este conceito e valores em relação a características de usuário que pode ser interpretado de uma forma diferente.

Ao comparar utilizando as comunidades (Figura 5.3), é obtido um grafo semelhante ao mostrado na métrica de *in-degree*, porém é adicionado o usuário que obteve um valor de *page rank* maior, mas com valor de *in-degree* baixo. Uma das principais características observadas da métrica *page rank*, com o conceito de comunidade, foi o fato do usuário menos privilegiado ("usuário#2606") estar na mesma comunidade do usuário que retweetou ("ddlloren"), existindo apenas este dois usuários com valores de *page rank* dentro da comunidade e o restante com valor menor que 0,001.

O usuário com maior valor de *page rank* foi, assim como na métrica *in-degree*, o "usuário#8653", mostrando que suas publicações foram espalhadas pela rede

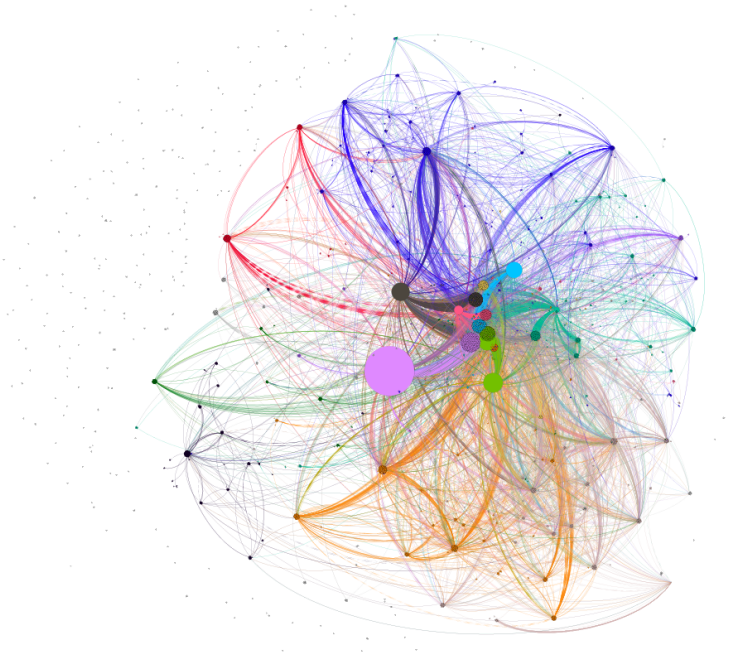

Figura 5.3: Classificação dos usuários baseados na métrica *Page Rank*

social utilizando as *hashtags* relacionadas ao evento, não precisando também de usuários específicos para difundir informações por toda rede social, uma vez que o usuário com maior *out-degree* de toda a rede não se encontra na mesma comunidade e não possuiu nenhuma interação direta com o mesmo (Figura **??**). Na próxima seção serão analisados aspectos relacionados ao usuário individualmente para determinar se os usuários considerados influentes permanecem com o mesmo status aplicando essas outras métricas.

5.1.2 Discussão

A partir das análises de todas as redes geradas, usuários coletados e métricas empregadas são percebidos alguns fatos que implicam na formalização das métricas propostas.

As métricas que mais respondem sobre um usuário influente neste nicho (racismo) são *in-degree* e *page rank*. Isso se deve pelo fato de que os usuários mais destacados são aqueles que mais possuem interações dentro da rede (*in-degree*) e também grau de importância (*page rank*). Analisando este último fator, é possível notar que através de um *retweet* de um usuário influente para outro, com menor influência, este último usuário torna-se mais vísivel para outros dentro da rede social.

Na Figura 5.4, o "Usuário#8653", representado pelo círculo mais afora do grafo, foi considerado o mais influente entre os usuários coletados e analisados pelas duas métricas de topologia (*in-degree* e *page rank*). Elas apontaram que este usuário recebe muitas interações, mas em contrapartida não interage com nenhum outro usuário e assim não compartilha sua importância com outros.

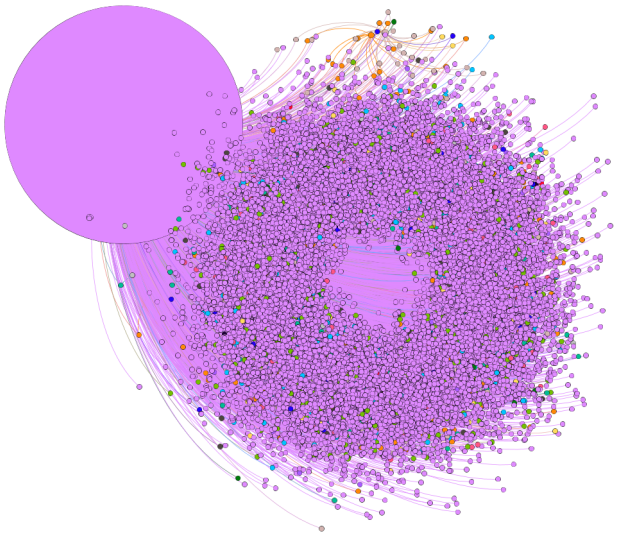

Figura 5.4: "Usuário#8653" considerado o mais influente utilizando os dois conjuntos de características

As métricas que caracterizam os usuários individualmente mostraram-se importantes quando são analisadas as questões de quantidade e qualidade de seguidores, frequência de publicações dos usuários dentro da rede social e quantidade de *retweets*.

A quantidade de seguidores refere-se a quantidade de outros usuários que são atingidos por publicações ou quaisquer atividades de um usuário principal, pois quanto mais seguidores mais visível este usuário principal se torna. A qualidade de seguidores reflete se seus seguidores possuem grandes interações e se são visíveis por muitos usuários dentro da rede social. A frequência de publicações pelo usuário mostra que as publicações realizadas dentro daquele evento coletado e utilizando *hashtags* que estão nos tópicos de tendência aumentam a visibilidade aquele *tweet* ganhará para outros usuários, e assim mais seguidores poderá conquistar.

Prova de tudo isso é o usuário mais "importante" dessa base. O "usuá-

rio#8653" ganhou notoriedade dentro da rede social após esse evento coletado, visto que sua publicação, no evento coletado, obteve mais de 37 mil *retweets*, mantendo-o popular e influente, conforme a Figura 5.5.

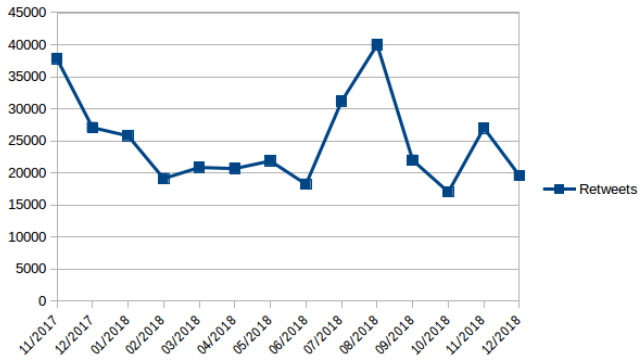

Figura 5.5: Análise de quantidade *retweets* do "Usuário#8653" por período

O gráfico aprensenta as quantidade de *retweets* adquiridas através dos meses de observação daquele usuário. É importante analisar que após o período no qual obteve o apice de importância (Novembro/2017), o usuário perdeu em quantidade de *retweets*. Porém, quando analisado a quantidade de seguidores, o "usuário#8653" obteve cinco vezes mais o que possuía quando não era considerado influente. Por consequência, passou a ser seguido por usuários com muitos seguidores, aumentando também a qualidade de seus seguidores.

Portanto, para este tipo de nicho fica claro que a quantidade de publicações não indica que este usuário possa se torna influente. O que mais determina, neste caso, é a quantidade de *retweets* que uma publicação possui. Além disso, as primeiras publicações se tornam as principais e ganham visibilidade por toda rede social. Entretanto, o fator usuários suspeitos ou *bots* não afetou diretamente na influência de um determinado usuário, visto que usuários considerados suspeitos não possuíam interações diretas.

5.2 Política

A base de dados denominada de Política foi gerada a partir da coleta de dados relacionadas a um tópico de tendência sobre o assunto. O grafo gerado possui mais de 43 mil usuários e mais 119 mil interações (*retweets* entre estes usuários). O assunto publicado e comentado é o julgamento de um ex-presidente da república, tratando-se de um tema polêmico.

Por se tratar de assunto no qual existem pessoas a favor e contra, foi decidido analisar primeiramente os dados, na forma de comunidades, como um grafo geral -

juntando os dois lados - e depois analisar separadamente os lados em duas bases (pró e contra prisão). O grafo resultante, apresentado na Figura 5.6, contém grandes comunidades para os dois tipos de movimentos. A parte mais superior do grafo indica usuários participantes do movimento pró prisão, contando com a presença de usuários importantes, se aplicada a medida *in-degree*, possuindo também mais usuários do que em relação ao movimento contra prisão.

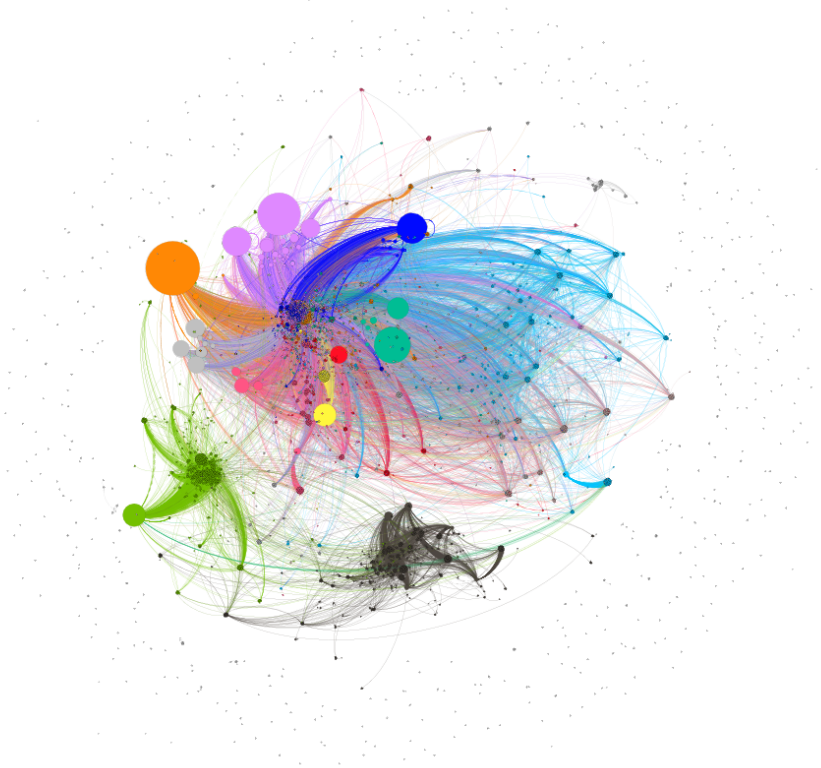

Figura 5.6: Classificação geral dos usuários (Pró e Contra) utilizando a métrica *In-Degree*

A lista dos cinquenta usuários mais importantes ou influentes dentro da rede gerada possui usuários dos dois tipos de movimentos, no entanto eles pouco interagem com o outro movimento. Essas interações iniciam quase sempre por usuários menos importantes de um movimento dentro da rede para usuários mais importante do outro movimento contrário. Esse tipo de comportamento é visto em ambos os lados.

As seções seguintes monstram a análise da base separadamente, através das métricas propostas. Logo, dois grafos são gerados e novamente comunidades são geradas a partir da separação realizada.

In-Degree

A primeira análise foi realizada em usuários que defendem o ex-presidente julgado (pró) utilizando as *hashtags*: #EleicaoSemLulaEFraude e #CadeAProva. Cada publicação que utilizou estas *hashtags* no período de uma semana foi coletada. Porém, nesta dissertação apenas usuários que obtiveram interações através de *retweets* foram considerados. Portanto, nesta primeira base foram encontrados um total de 9.593 usuários possuindo um total de 14.807 interações, os quais são analisados na forma de comunidades como na base anterior e apresentados na Figura 5.7.

Figura 5.7: Classificação dos usuários "Pró" baseados na métrica *In-Degree*

A Figura mostra os usuários classificados no aspecto da métrica *in-degree*, onde poucos obtiveram destaques, possuindo valores próximos entre os usuários analisados. Apenas o primeiro usuário ("usuário#7961"), de cor lilás e maior nó dentro da rede, obteve uma importância diferente dos demais com 2.366 interações. O segundo usuário tem uma diferença de mais mil para o primeiro como mostra a Tabela 5.4.

Tabela 5.4: Classificação dos usuários "Pró" baseados na métrica *in-degree*

	ID usuário	***In-degree***
1	usuário#7961	2.366
2	usuário#7249	806
3	usuário#4529	710
4	usuário#3129	414
5	usuário#1349	351
6	usuário#1483	342
7	usuário#1540	320
8	usuário#4707	273
9	usuário#7604	272
10	usuário#6250	271

Conforme a Tabela, os usuários que representam o movimento pró possuem um equilíbrio nos valores das métricas de *in-degree*, com a única exceção sendo o primeiro usuário, que recebeu *retweets*.a Um outro ponto importante encontrado é que o usuário "comunacritico" é o único dentro de uma comunidade específica que possuiu um alto valor de *in-degree*, o que é gerado a partir de seus seguidores, sendo comprovado na próxima métrica analisada (*out-degree*).

Outro fator é a quantidade de usuários que não são do país, mas sim da América Latina, tais como, "usuário#7249", "usuário#4529", "usuário#3129", "usuário#1483", entre outros. Isso se deve a quantidade seguidores que os usuários possuem, uma vez que eles espalham as informações dentro da rede através de *retweets* e menções a usuários importantes, fazendo com que outros se interessem pelo conteúdo e assim começem a repassá-los. Também existe o fato de políticos estarem entre os mais importante nesta métrica, possuindo grande influência sobre os seus seguidores.

Para analisar o movimento Contra o julgamento do ex-presidente foram utilizadas as *hashtags*: #MoluscoNaCadeia e #CondenaTRF4. Assim como no movimento Pró, cada publicação utilizou estas *hashtags* no período de uma semana em que foram coletadas, mas, nesta dissertação, apenas usuários que obtiveram interações através de *retweets* foram considerados. Portanto, nesta base foram encontrados um total de 34.665 usuários com um total de 89.662 interações, um significativo aumento em relação ao movimento anterior. Como na base anterior, os dados são analisados na forma de comunidades e apresentados na Figura 5.8.

Assim como base contra o julgamento, o movimento pró apresenta uma grande diferença entre o primeiro e o segundo usuário para a métrica *in-degree*, sendo tal diferença quase o dobro para o segundo, conforme a Tabela 5.5.

Porém, um ponto que difere é quantidade de comunidades encontradas. Os

Figura 5.8: Classificação dos usuários "Contra" baseados na métrica *In-Degree*

Tabela 5.5: Classificação dos usuários "Contra" baseados na métrica *in-degree*

	ID usuário	***In-degree***
1	usuário#7745	4.490
2	usuário#8407	2.953
3	usuário#8713	2.407
4	usuário#9419	2.350
5	usuário#8300	1.869
6	usuário#1750	1.770
7	usuário#7098	1.686
8	usuário#8554	1.632
9	usuário#2571	1.545
10	usuário#2884	1.506

usuários neste movimento (contra) participam de diferentes e diversas comunidades, logo nenhum dos primeiros na métrica analisada estão dentro da mesma

comunidade. Isso ocorre pela grande quantidade de *retweets* distribuídos entre os usuários. Entre os dez primeiros representados, todos são de comunidades diferentes. Isso se deve pelo fato desta rede possuir mais usuários e as *hashtags* coletadas permaneceram por três dias seguidos nos principais tópicos de tendências, fazendo com que ocorram mais interações que no movimento anterior.

Out-degree

Esta métrica revelou existir uma grande comunidade de usuários que retweetaram diversos outros. Essa comunidade específica foi responsável por distribuir *retweets* para todas as outras comunidades dentro da rede, inclusive para aquelas que possuíram maior *in-degree*.

Os usuários classificados nesta métrica, e que estão na base de dados, frequentemente *retweetaram* tipos específicos de usuários, como forma de difundir as publicações dos mesmos e torná-los ainda mais influente entre seus seguidores. Este tipo de usuário é conhecido como difusor de informação, muita das vezes podendo ser um usuário falso ou os chamados *bots* programados para realizar *retweets* ou publicações que mencionem aquele determinado perfil.

Quando se trata do movimento contra nesta mesma métrica, os valores tornam-se muito grandes, pois diversos usuários atuaram de forma bastante participativa dentro dos tópicos de tendência. Com isso, a quantidade de usuários que retweetaram outros e também suas publicações realizadas em todo o evento coletado é diferente quando comparado com a base do movimento pró. De fato, não foram somente os primeiros mais bem classificados nesta base que realizaram interações, mas sim mais de cem usuários ativamente interagindo, sendo a comunidade mais superior de cor lilás (Figura 5.2) com a maior quantidade de usuários que difundiram publicações dentro da rede social.

A comunidade mais destacada nesta métrica possui usuários comuns, nos quais utilizam do mesmo artifício para ganhar popularidade dentro da rede social: retweetar usuários mais influente para que possam ganhar novos seguidores e assim suas publicações se tornarem ainda mais visíveis a outros, tornando-sepontes para outros usuários mais influentes, como observado na próxima métrica analisada.

Betweenness e Closeness

Nas duas métricas envolvendo a centralidade dos usuários quanto a distância para outros, *betweenness*, a base envolvendo o movimento pró obteve a mesma classificação referente a Tabela 5.4, onde os principais usuários são pontes para os demais, significando que um usuário não importante ao retweetar um certo usuário mais influente é provável que o primeiro tenha sido conhecidos por outros

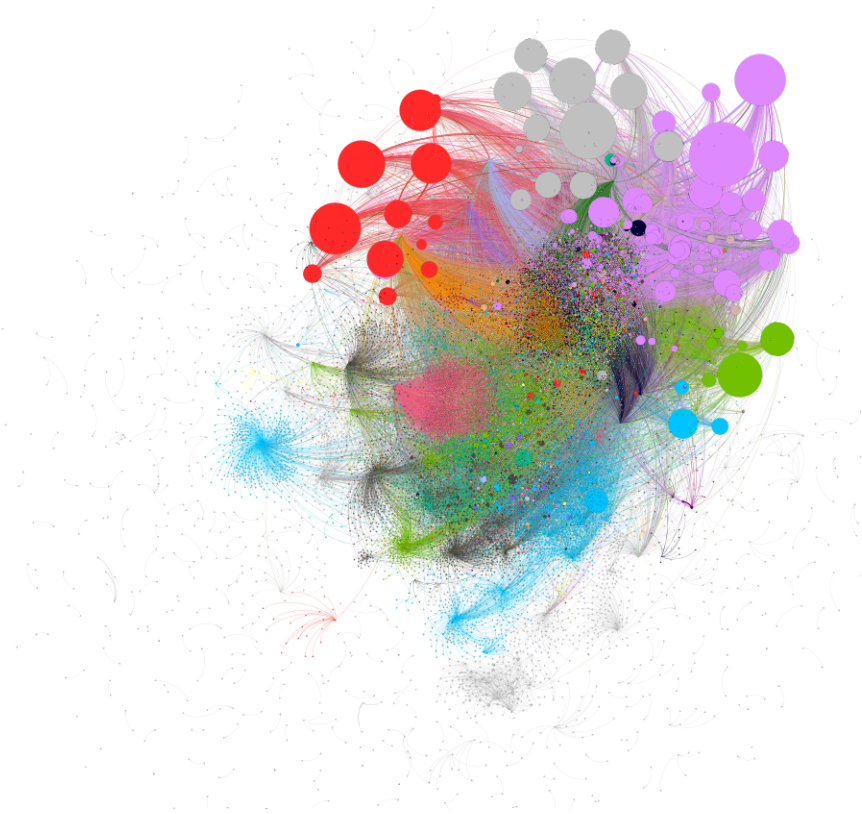

Figura 5.9: Classificação dos usuários “Contra” baseados na métrica *Out-Degree*

usuários influentes dentro da rede ou simplesmente seguidores de seguidores.

Neste tipo de métrica, se tratando especificamente desta base, os usuários tem diversas informações passando por seus seguidores, o que acaba por formar várias comunidades, visto que opiniões parecidas vão criando comunidades.

Os dois movimentos possuem como usuários centrais e principais aqueles que tiveram publicações mais polêmicas. Dentro desse conjunto, estão usuários famosos, perfis de comédia e movimentos sociais. Todos estes geralmente possuem um grande número de seguidores ativos, que acompanham toda atualização realizada e assim difundem as mensagens para outros perfis.

Page Rank

A métrica de *Page Rank* identificou os usuários considerados influentes dentro do movimento pró. Novamente, existem usuários que na medida *in-degree* não estavam entre os principais, mas se tratando de *page rank* se tornaram influentes. Usuários com alto *in-degree* retweetaram aqueles que possuíram um baixo valor nesta métrica, como pode se observar na Tabela 5.6 comparando os valores das duas métricas.

Tabela 5.6: Classificação dos usuários "Pró" baseados na métrica *Page Rank*

	ID usuário	***Page Rank***	***In-degree***
1	usuário#7961	0,052	2,366
2	usuário#7249	0,021	806
3	usuário#1111	0,015	265
4	usuário#4707	0,013	273
5	usuário#4529	0,012	710
6	usuário#6250	0,010	271
7	usuário#2222	0,007	129
8	usuário#7604	0,007	272
9	usuário#3129	0,007	414
10	usuário#1483	0,007	342

Observando a Tabela, percebe-se que apareceram novos usuários. Vale destacar que a definição da métrica é ponderar os usuários que possuem grandes interações (*in-degree*) com usuários menos influentes como forma de distribuir bem a importância entre todos os nós analisados. Desta forma, usuários antes menos influentes se tornam mais importantes dentro da rede, mass não necessariamente influente. O grafo que representa esses usuários é representado na Figura 5.7.

Na Tabela, os usuários “usuário#1111” e “usuário#2222”, que antes não estavam entre os dez primeiros na primeira métrica analisada, agora estão entre os mais importantes. Isso acontece pelo fato de que o usuário mais importante neste caso, o “usuário#7249”, retweetou o usuário “usuário#1111”, fazendo com que este último tivesse uma importância parecida. Outro fator é que os dois não estão dentro da mesma comunidade, fazendo que o “usuário#1111” ganhe ainda mais destaque dentro do evento coletado.

Outro caso acontece com o usuário “usuário#2222”, que também não estava entre os dez primeiros na métrica *in-degree*, mas como foi retweetado pelo usuário “usuário#6250”. Como este último se trata de um usuário político, logo tem uma influência sobre seus seguidores. Isso faz com que o “usuário#2222” se torne ainda mais visível dentro da rede. Os dois usuários estão na mesma comunidade analisada.

Utilizando a métrica *page rank* para a segunda base são obtidos valores praticamente iguais ao encontrados na métrica de *in-degree*, não havendo usuários que ganharam mais destaque ao realizar uma interação com um usuário mais influente dentro da rede.

Ao analisar em conjunto os dois tipos de características, observa-se que existem diversos tipos de usuários influenciadores nas duas bases de dados (pró e contra). Um dado importante é o tipo de perfil que realiza publicações sobre o tema, desde perfis de comediantes a portais de notícias. São perfis mais consolidados em termos de seguidores, tanto em quantidade quanto em qualidade, pois são seguidos por outros usuários importantes do mesmo nicho, o que explica quando eles estão dentro de uma mesma comunidade.

O “usuário#7961” (Figura 5.10) foi considerado o mais influente também quando o seu perfil é analisado individualmente, por ser um usuário que realiza críticas neste tipo de evento, possui muitos seguidores (mais de 30 mil) e também possui boas interações. Um ponto negativo em consideração a este usuário é que o mesmo nem sempre possui um grande quantidade de *retweets* em suas publicações. Este fato se deve à quantidade de publicações que são realizadas pelo usuário dentro da rede social, fazendo que seus seguidores não sigam totalmente todas as publicações fielmente. Fica observado também que para este tipo de evento, os usuários possuem um mesmo padrão quanto a periodicidade de publicaçs dentro da rede social.

O “usuário#7745” (Figura 5.11) foi considerado influente dentro da rede analisada em perfis que são de movimento contrário. É um usuário conhecido pelo público por realizar sarcasmos, ironia e comédia, obtendo assim uma grande quantidade de retweets por seus seguidores e não seguidores. Com isso, o perfil tem mais 100 mil seguidores até o momento da coleta de seus dados. Como pode ser visto na Figura, o “usuário#7745’ se conecta diretamente com outros usuários

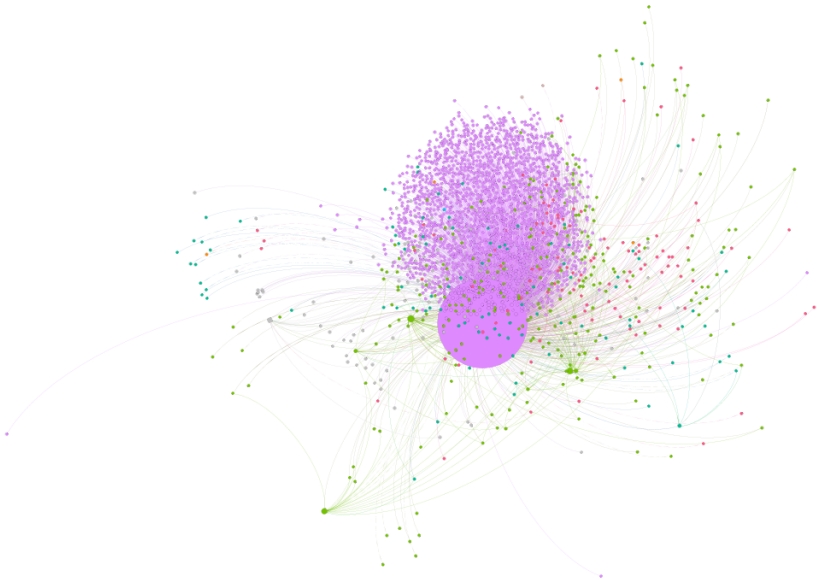

Figura 5.10: Análise de quantidade *retweets* do “Usuário#7961” por período

considerados influentes dentro da rede social através de suas publicações durante o evento. Esses perfis se assemelham pelo tipo de publicações que realizam, por isso encontram-se dentro da mesma comunidade.

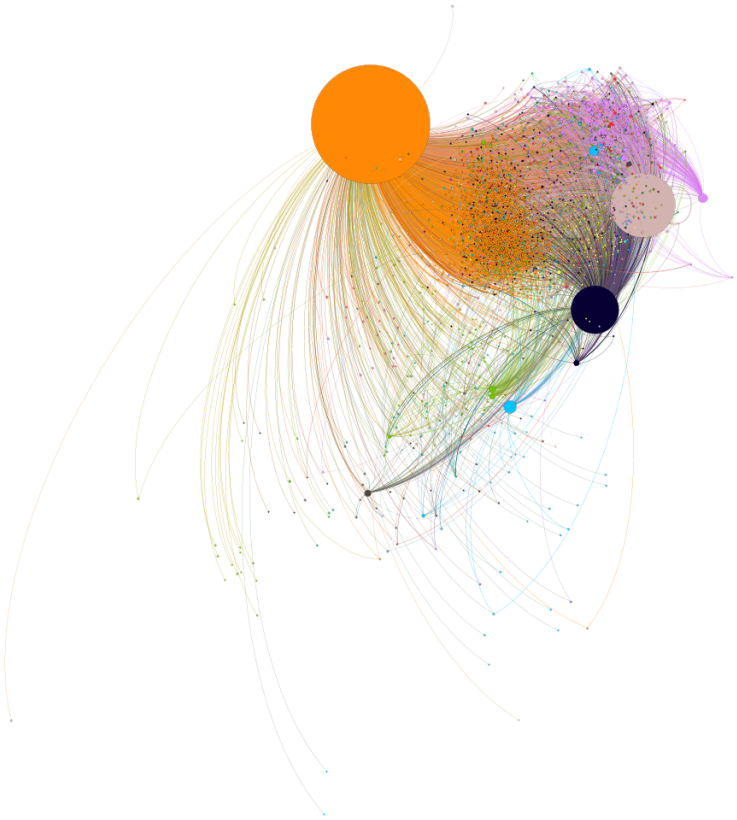

Figura 5.11: Análise de quantidade *retweets* do "Usuário#7745" por período

Sendo assim, pode-se afirmar que para este tipo de evento coletado (política), usuários que costumam ser influentes durante seu período de vida na rede social continuam sendo influentes para todo e qualquer evento desta natureza, pois a quantidade de seguidores dos mesmos influencia de forma direta tanto na sua rede local quanto global. Estes usuários influentes abrangem e conquistam outros seguidores que, na maioria das vezes, não são experientes ou possuem opinião sobre o tema discutido.

Estes perfis influentes dentro deste tipo evento possuem particularidades que são observadas através da análise dos dois conjuntos propostos. São perfis que possuem um engajamento relacionado ao evento, onde mais da metade de suas publicações são redirecionadas ao tema coletado. Possuem uma grande quantidade de seguidores e são seguidos por seguidores também influentes, o que

resulta na divisão entre importância entre dois usuários (*page rank*). Além disso, são perfis já consolidados dentro da rede social, ou seja, não ganharam seguidores e influência através de poucas publicações, mas sim com o passar do tempo, realizando diversos *tweets* - outra característica encontrada dentro desta base de dados. Esses usuários que realizam muitas publicações diárias em comparação com o primeiro evento analisado (racismo).

5.3 Discussões

Conforme apresentado no Capítulo, as métricas empregadas respondem diversas questões que podem ser criadas através dos dados coletados. Nesta dissertação foram propostas duas questões de pesquisa e para respondê-las foram aplicadas métricas nas quais utilizam dois tipos diferente de análise: topologia e usuário. Uma vez que na literatura autores propõem abordagens utilizando apenas um dos tipos, fazendo que os dados não represente sempre 100% fielmente o que são no mundo real.

Durante este estudo foi observado um pouco do comportamento deste usuários (antes, durante e depois), mas sempre comparando com o que foi observado durante o tópico de tendência ser analisado, ou seja, o antes e depois foram observados depois da coleta e análise dentro do tópico. Com isso, ficou notado que muitos usuários ganham notoriedade dentro da rede social não só por *tweets* que expressam opiniões, mas como também pela quantidade de seguidores, a qualidade de seus seguidores ou pelo fato de sua publicação estar em destaque dentro da rede social.

Para a base de dados 1, fica esclarecido que usuários menos conhecidos podem ganhar importância de acordo com a quantidade publicações dentro daquele tópico coletado. Um exemplo é o perfil mais retweetado desta base, que após o evento aumentou a sua quantidade de seguidores em mais 10 mil usuários. Porém, utilizando apenas uma publicação para em todos os dias que foram coletados. Mas isso não significa que antes do evento o mesmo não possuía uma grande popularidade, mas sim que o evento causou um grande aumento na quantidade de seus seguidores.

A base de dados 2 utiliza um tema bastante corriqueiro (política). Para este tópico coletado, os usuários tenderam a ser mais participativos quando se analisa o número de interações e publicações. Este tipo evento tendem a ter um diversificado tipo de usuário, que por sua vez criam interações com outros sem possuir necessariamente um relacionamento criado dentro da rede social. Outro ponto esclarecido é a quantidade de usuários que servem como difusores de informação para ganhar notoriedade e futuramente suas publicações abrangerem mais usuários. Os usuários considerados mais influentes neste tipo de evento são

aqueles que possuem uma alta quantidade de seguidores e possuem também suas publicações com altos índices de *retweet*.

Portanto, para realizar análise de influência e difusão de informação em redes sociais, é preciso não somente analisar métricas que já foram criadas, mas também analisar cenários diferentes, tipos de usuários, comportamento, importância do cenário escolhido dentro do mundo real, dentre outros. Contudo, o próximo Capítulo responde as questões propostas analisando estes tipos de fatores citados.

Capítulo 6

Conclusões e Trabalhos Futuros

Nesta dissertação foi proposto um estudo sobre métricas que envolvem dois tipos de características (topológicas e de usuários) para determinar influência e difusão de informação dentro de redes sociais, neste caso, o Twitter. Analisando as métricas em conjunto utilizando eventos provenientes de tópicos de tendências mais comentados nos dias ou semana.

Para isso, foi implementada uma metodologia que começa na fase de coleta de dados, onde é responsável por coletar publicações, *retweets* recebidos, nome do usuário, seguindo e seguidores. Após essa fase, foi implementado um script responsável por coletar amizade dos usuários, ou seja, se ao retweetar um certo usuário, qual o nível de relacionamento os dois possuem dentro da rede social. Neste caso são utilizadas três tipos de relacionamentos, como são descritos abaixo.

- Seguindo: onde o usuário que retweeta a publicação apenas segue o receptor,
- Seguidor: no qual receptor segue o usuário que o retweetou,
- Mútuo: há uma reciprocidade no relacionamento seguindo/seguidor

A construção deste script foi requisitada, pois os trabalhos encontrados na literatura, não avaliam de forma profunda a questão da amizade entre usuários para realizar o estudo de influência em redes sociais. E para esta dissertação ficou notado a importância de haver este estudo utilizando também este conceito de relacionamento entre usuários, pois essa informação nos traz bastante fundamentação para analisar se o usuário pode ser considerado influente ou não e se o mesmo difunde informações para toda rede social ou apenas para sua rede mais local, ou seja, seus seguidores.

As bases coletadas são provenientes de dois assuntos distintos mas que se encaixam no tema de polêmicos, pois tratam de caso de racismo e política. O

objetivo foi analisar como usuários se comportam nestes dois assuntos e se caso eles conseguem ter opiniões, influência ou importância nas duas bases coletadas, a forma de interação que foi utilizada entre os usuários nesta dissertação foi o *retweet*, onde o ato de retweetar uma publicação podem dizer muito sobre o usuário analisado.

Com estas informações os usuários foram dispostos em forma de grafos para assim poder computar suas métricas envolvendo as duas características citadas. Estes usuários são dispostos em comunidades como forma de melhor representar e também avaliar em contexto mais realista os relacionamentos dos mesmos dentro da rede social. Logo após, foram aplicados as métricas dentro de cada rede ou evento coletado, métricas essas que revelaram resultados diferentes observando o mesmo ou contexto diferentes.

A análise sobre as métricas utilizadas tiveram o objetivo de responder à duas questões de pesquisa elaboradas no Capítulo 1, nas quais buscam identificar usuários influente e difusão de informação na rede social. A primeira pergunta: "É possível identificar usuários influentes em nichos ou temas específicos em redes sociais?", trata-se de afirmar se há possibilidade de encontrar usuários influentes nos dois nichos analisados: racismo e política.

Para esta pergunta, a primeira base analisada identificou usuários que são possíveis de serem influentes dentro da rede social, pois combinando as métricas topológicas e de usuário, foram eles que se sobressaíram em relação a outros usuários menos relevantes. Entretanto, ao analisar histórico de publicações sobre a quantidade de *retweets* gerados, consegue-se observar que estes usuários não são influentes dentro da rede social, pois em outras publicações fora deste nicho os mesmos não possuem valores tão expressivos.

Diferentemente do que acontece com a segunda base coletada, na qual se trata de política, onde pode-se observar que usuários nos quais obtiveram os maiores valores nas métricas possuem opiniões importante dentro daquele nicho (política) específico, pois são usuários que tratam diretamente com aquele público, por isso possuem uma facilidade maior de interação com seus seguidores e não seguidores.

Em relação a segunda questão de pesquisa: "As métricas encontradas hoje na literatura conseguem representar fielmente influência e difusão de informação em redes sociais?", na qual se trata analisar as métricas (topológicas e usuário) escolhidas para esta dissertação e afirmar positivamente ou negativamente quanto à sua relevância. As métricas analisadas individualmente possuem diferentes propósitos, seja para analisar e avaliar o quanto central um usuário foi durante um tópico de tendência específico.

Entretanto que ficou observado utilizando nas bases de dados é que ao usar apenas uma métrica para avaliar e representar fielmente aquele usuário como influente é errado, pois as métricas só conseguem responder para aquele período

de tempo coletado e não considera o passado e o tempo de vida daquele perfil na rede social. Este fator ficou observado para duas bases analisadas, pois ao avaliar apenas características topológicas, alguns usuários conseguiram obter altos valores para as métricas definidas, por exemplo utilizando o *in-degree* é possível afirmar que aquele usuário pode ser um influente dentro da rede social. Contudo ao analisar suas características mais locais utilizando informações sobre seguidores, número de publicações dentro do evento e os seus *retweets* gerados, fica claro que aquele mesmo usuário não possui uma importância tão grande assim dentro da rede social ou vice-versa.

Uma métrica que ao ser analisada pode declinar qualquer métricas analisadas individualmente é a *out-degree*, por se tratar de interações realizadas por um certo usuário diversas vezes esse usuário possui comportamento estranho dentro da rede social, ou seja, retweetar usuários sem critérios como forma de conseguir seguidores, seguir usuários como forma também de obter mais seguidores, ou tratando-se de perfil contratado para interagir somente com um perfil específico e partir disso espalhar suas informações por toda rede social e torna-se um usuário influente. Portanto, os usuários que tem um alto valor nesta métrica (*out-degree*) são considerados como difusores de informação dentro da rede social, quanto para afirmar a questão de influência de um usuário é preciso combinar métricas globais (topológicas) e locais (usuários), para assim se aproximar de um resultado mais cabível.

6.1 Trabalhos Futuros

No entanto, para obter-se resultados mais aprimorados para este assunto de influência e difusão de informação são necessários estudos mais aprofundados tanto nas métricas quanto nas bases coletadas, pois o que se observa o maior problema não nas métricas já existentes, mas sim na sua aplicação dentro das bases de dados coletadas, para isso são identificados trabalhos futuros que podem melhorar estes aspectos.

- Criar abordagens mais diversificadas utilizando dados que API da rede provem com a finalidade torná-las métricas ou aprimorar as existentes.
- Realizar estudos analisando o conceito de *bot* e seu comportamento dentro das redes de influência.
- Manipular mais bases de dados como forma de diferenciar ou identificar mais aspectos que podem separar nichos extremamente distintos, por exemplo, economia e esportes.

Referências Bibliográficas

[1] Al-garadi, M. A., Varathan, K. D., and Ravana, S. D. (2017). Identification of influential spreaders in online social networks using interaction weighted k-core decomposition method. *Physica A: Statistical Mechanics and its Applications*, 468:278–288.

[2] Al-Garadi, M. A., Varathan, K. D., Ravana, S. D., Ahmed, E., and Chang, V. (2016). Identifying the influential spreaders in multilayer interactions of online social networks. *Journal of Intelligent & Fuzzy Systems*, 31(5):2721–2735.

[3] Al-Garadi, M. A., Varathan, K. D., Ravana, S. D., Ahmed, E., Mujtaba, G., Khan, M. U. S., and Khan, S. U. (2018). Analysis of online social network connections for identification of influential users: Survey and open research issues. *ACM Computing Surveys (CSUR)*, 51(1):16.

[4] Aleahmad, A., Karisani, P., Rahgozar, M., and Oroumchian, F. (2016). Olfinder: Finding opinion leaders in online social networks. *Journal of Information Science*, 42(5):659–674.

[5] Alp, Z. Z. and Öğüdücü, Ş. G. (2018). Identifying topical influencers on twitter based on user behavior and network topology. *Knowledge-Based Systems*, 141:211–221.

[6] Asadi, M. and Agah, A. (2017). Characterizing user influence within twitter. In *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, pages 122–132. Springer.

[7] Beveridge, A. and Shan, J. (2016). Network of thrones. *Math Horizons*, 23(4):18–22.

[8] Bigonha, C., Cardoso, T. N., Moro, M. M., Gonçalves, M. A., and Almeida, V. A. (2012). Sentiment-based influence detection on twitter. *Journal of the Brazilian Computer Society*, 18(3):169–183.

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

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

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

[12] Cerón-Guzmán, J. A. and León, E. (2015). Detecting social spammers in colombia 2014 presidential election. In *Mexican International Conference on Artificial Intelligence*, pages 121–141. Springer.

[13] Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P. K., et al. (2010). Measuring user influence in twitter: The million follower fallacy. *Icwsm*, 10(10-17):30.

[14] Chai, W., Xu, W., Zuo, M., and Wen, X. (2013). Acqr: A novel framework to identify and predict influential users in micro-blogging. In *Pacis*, page 20.

[15] De Domenico, M., Lima, A., Mougel, P., and Musolesi, M. (2013). The anatomy of a scientific rumor. *Scientific reports*, 3:2980.

[16] del Fresno Garcia, M., Daly, A. J., and Segado Sanchez-Cabezudo, S. (2016). Identifying the new influences in the internet era: Social media and social network analysis. *Revista Española de Investigaciones Sociológicas*, (153).

[17] Díaz-Beristain, Y. A., Cruz-Ramírez, N., et al. (2016). Retweet influence on user popularity over time: An empirical study. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 38–48. Springer.

[18] Dubois, E. and Gaffney, D. (2014). The multiple facets of influence: Identifying political influentials and opinion leaders on twitter. *American Behavioral Scientist*, 58(10):1260–1277.

[19] Easley, D. and Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.

[20] Ellison, N. B. et al. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230.

[21] Farahani, H. S., Bagheri, A., and Saraf, E. H. K. M. (2017). Characterizing behavior of topical authorities in twitter. In *Innovative Mechanisms for Industry Applications (ICIMIA), 2017 International Conference on*, pages 581–586. IEEE.

[22] Ferreira, A. B. d. H. (2004). Novo dicionário aurélio da língua portuguesa. In *Novo dicionário Aurélio da língua portuguesa*.

[23] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

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

[25] Katz, E. and Lazarsfeld, P. F. (1955). Personal influence: the part played by people in the flow of mass communications.

[26] Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., and Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature physics*, 6(11):888.

[27] Lahuerta-Otero, E. and Cordero-Gutiérrez, R. (2016). Looking for the perfect tweet. the use of data mining techniques to find influencers on twitter. *Computers in Human Behavior*, 64:575–583.

[28] Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., and Choudhary, A. (2011). Twitter trending topic classification. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 251–258. IEEE.

[29] Leskovec, J. and Krevl, A. (2014). SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data.

[30] Li, J., Peng, W., Li, T., Sun, T., Li, Q., and Xu, J. (2014). Social network user influence sense-making and dynamics prediction. *Expert Systems with Applications*, 41(11):5115–5124.

[31] Liu, Y., Tang, M., Zhou, T., and Do, Y. (2015). Core-like groups result in invalidation of identifying super-spreader by k-shell decomposition. *Scientific reports*, 5:9602.

[32] Lü, L., Chen, D., Ren, X.-L., Zhang, Q.-M., Zhang, Y.-C., and Zhou, T. (2016). Vital nodes identification in complex networks. *Physics Reports*, 650:1–63.

[33] Mei, Y., Zhong, Y., and Yang, J. (2015). Finding and analyzing principal features for measuring user influence on twitter. In *Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on*, pages 478–486. IEEE.

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

[35] Nagmoti, R., Teredesai, A., and De Cock, M. (2010). Ranking approaches for microblog search. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 153–157. IEEE Computer Society.

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

[37] Pal, A. and Counts, S. (2011). Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 45–54. ACM.

[38] Pei, S., Muchnik, L., Andrade Jr, J. S., Zheng, Z., and Makse, H. A. (2014). Searching for superspreaders of information in real-world social media. *Scientific reports*, 4:5547.

[39] Räbiger, S. and Spiliopoulou, M. (2015). A framework for validating the merit of properties that predict the influence of a twitter user. *Expert Systems with Applications*, 42(5):2824–2834.

[40] Riquelme, F. and González-Cantergiani, P. (2016). Measuring user influence on twitter: A survey. *Information Processing & Management*, 52(5):949–975.

[41] Rochat, Y. (2009). Closeness centrality extended to unconnected graphs: The harmonic centrality index. In *ASNA*, number EPFL-CONF-200525.

[42] Statista (2017). Leading social networks worldwide as of april 2017. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[43] Subramani, M. R. and Rajagopalan, B. (2003). Knowledge-sharing and influence in online social networks via viral marketing. *Communications of the ACM*, 46(12):300–307.

[44] Sun, Q., Wang, N., Zhou, Y., and Luo, Z. (2016). Identification of influential online social network users based on multi-features. *International Journal of Pattern Recognition and Artificial Intelligence*, 30(06):1659015.

[45] Wellman, B. (1996). For a social network analysis of computer networks: a sociological perspective on collaborative work and virtual community. In *Proceedings of the 1996 ACM SIGCPR/SIGMIS conference on Computer personnel research*, pages 1–11. ACM.

[46] Wen, S., Jiang, J., Xiang, Y., Yu, S., Zhou, W., and Jia, W. (2014). To shut them up or to clarify: Restraining the spread of rumors in online social networks. *IEEE Transactions on Parallel and Distributed Systems*, 25(12):3306–3316.

[47] Weng, L., Flammini, A., Vespignani, A., and Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific reports*, 2:335.

[48] Weng, L., Menczer, F., and Ahn, Y.-Y. (2013). Virality prediction and community structure in social networks. *Scientific reports*, 3:2522.

[49] Wives, L. K. (2013). *Descobrindo eventos locais utilizando análise de séries temporais nos dados do Twitter*. PhD thesis, UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL.

[50] Yang, L., Qiao, Y., Liu, Z., Ma, J., and Li, X. (2018). Identifying opinion leader nodes in online social networks with a new closeness evaluation algorithm. *Soft Computing*, 22(2):453–464.

[51] Zafarani, R., Abbasi, M. A., and Liu, H. (2014). *Social media mining: an introduction*. Cambridge University Press.

[52] Zamparas, V., Kanavos, A., and Makris, C. (2015). Real time analytics for measuring user influence on twitter. In *Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on*, pages 591–597. IEEE.

[53] Zarrella, D. (2009). The science of retweets. *Retrieved December*, 15:2009.

[54] Zhang, X., Li, Z., Zhu, S., and Liang, W. (2016a). Detecting spam and promoting campaigns in twitter. *ACM Transactions on the Web (TWEB)*, 10(1):4.

[55] Zhang, Z.-K., Liu, C., Zhan, X.-X., Lu, X., Zhang, C.-X., and Zhang, Y.-C. (2016b). Dynamics of information diffusion and its applications on complex networks. *Physics Reports*, 651:1–34.

[56] Zhao, K., Greer, G. E., Yen, J., Mitra, P., and Portier, K. (2015). Leader identification in an online health community for cancer survivors: a social

network-based classification approach. *Information Systems and e-Business Management*, 13(4):629–645.

[57] Zhu, H., Huang, C., and Li, H. (2014). Mppm: Malware propagation and prevention model in online sns. In *Communications Workshops (ICC), 2014 IEEE International Conference on*, pages 682–687. IEEE.

[58] Zhuang, K., Shen, H., and Zhang, H. (2017). User spread influence measurement in microblog. *Multimedia Tools and Applications*, 76(3):3169–3185.