Estrutura e dinâmica de redes de informação

Luís Fernando Dorelli de Abreu

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Luís Fernando Dorelli de Abreu

Estrutura e dinâmica de redes de informação

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Francisco Aparecido Rodrigues

USP – São Carlos
Julho de 2016

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

A634e Abreu, Luís Fernando Dorelli de
 Estrutura e dinâmica de redes de informação /
 Luís Fernando Dorelli de Abreu; orientador Francisco
 Aparecido Rodrigues. - São Carlos - SP, 2016.
 75 p.

 Dissertação (Mestrado - Programa de Pós-Graduação
 em Ciências de Computação e Matemática Computacional)
 - Instituto de Ciências Matemáticas e de Computação,
 Universidade de São Paulo, 2016.

 1. Redes complexas. 2. Grafos. 3. Redes Sociais.
 I. Rodrigues, Francisco Aparecido, orient. II.
 Título.

Luís Fernando Dorelli de Abreu

Structure and dynamics of information networks

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Francisco Aparecido Rodrigues

USP – São Carlos
July 2016

Dedico esse trabalho a curiosidade, ao desejo de descobrir e a coragem de tentar.

AGRADECIMENTOS

Começo os agradecimentos citando meus amigos em geral, pois qualquer conquista, por mais excitante que seja, não vale de nada sem amigos para comemorar. Essa dissertação não é diferente.

Meu agradecimento principal é direcionado ao meu orientador, Francisco, que foi paciente e me deu o suporte necessário para realizar esta pesquisa e que me motivou a realizar um mestrado em primeiro lugar.

Todos os membros do Grupo de Estudos para Maratona de Programação do ICMC merecem ser citados nesse trabalho pelas inúmeras experiências positivas nos últimos anos, sem as quais eu com certeza teria desanimado ao longo do caminho.

Agradeço também a minha família. Um agradecimento especial a minha avó Leni, que com 84 anos continua firme e forte e sempre dizia que não iria viver para me ver graduado, e a minha mãe e irmã, que tenho visto tão pouco tempo nos últimos anos mas que são meu chão.

Agradecimentos especiais são direcionados à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa concedida durante o início do curso de Mestrado e à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), processo 2014/12335-0, pela bolsa concedida durante o restante do curso.

“Alegações extraordinárias exigem evidências extraordinárias.”

(Carl Sagan)

RESUMO

DORELLI, L. F. A.. **Estrutura e dinâmica de redes de informação**. 2016. 75 f. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

O aumento na disponibilidade de dados referentes a interação entre pessoas *online* tornou possível o estudo o processo de propagação de informações em redes sociais com volumes de dado antes jamais pensados. Neste trabalho, utilizamos dados do site de *micro-blogging Twitter* juntamente com conceitos de redes complexas para entender, caracterizar e classificar processos de difusão de informação observados nessa plataforma e em redes sociais em *geral*. Apresentamos importantes medidas para caracterização de cascatas de informação, bem como algoritmos eficientes para o seu cálculo. Com o auxílio dessas, mostramos que é possível quantificar a influência da rede social no processo de propagação de informação. Em seguida, constatamos que a informação tende a propagar por caminhos mínimos nessa rede. Por fim, mostramos que é possível utilizar apenas a topologia da rede social, sem nenhuma informação semântica, para agrupar tópicos, e que a topologia da rede social é fortemente influenciada pelos assuntos falados nela. Apesar de nosso trabalho possuir como base um único *dataset*, os métodos e medidas desenvolvidos são gerais e podem ser aplicados a qualquer processo de difusão de informação e a qualquer rede complexa.

Palavras-chave: Redes complexas, Grafos, Redes Sociais.

ABSTRACT

DORELLI, L. F. A.. **Estrutura e dinâmica de redes de informação**. 2016. 75 f. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

The raise in the availability of data regarding interactions between people online has opened new doors to study the process of information diffusion in social networks. In this present work, we make use of the data from the micro-blogging website *Twitter* along with complex networks concepts to understand, characterize and classify information diffusion processes observed in this platform and in social networks in general. We present important measures to characterize information cascades and efficient algorithms to calculate them. With the help of these measures, we show that it is possible to quantify the influence of the social network in the process of information diffusion. After that, we show that information does tend to travel along shortest paths on *Twitter*. Finally, we show that the topology of the social network, without any extra semantic information, can be used to aggregate topics, and that such topology is highly influenced by the topics being discussed on it. Although we work with only a single dataset, our methods and measures developed are general and can be applied to any process of information diffusion and any complex network.

Key-words: Complex networks, Graphs, Social networks.

LISTA DE ILUSTRAÇÕES

Figura 1 – Possíveis formas de conexão no <i>Twitter</i>	28
Figura 2 – Exemplo de pacote <i>json</i> contendo um <i>tweet</i>	29
Figura 3 – Um grafo não-directionado e sua matriz de adjacência	30
Figura 4 – Ilustração esquemática de três situações onde o coeficiente de aglomeração apresenta diferentes valores.	31
Figura 5 – (a) Rede aleatória com $N = 300$ e $p = 0.1$ (b) Distribuição do grau para uma rede ER com 10000 vértices e $p = 0.5$	34
Figura 6 – Redes geradas pelo modelo WS com $K = 4$ e $N = 25$ quando (a) $p = 0.0$, (b) $p = 0.2$, (c) $p = 0.5$ e (d) $p = 0.8$. Quando $p \rightarrow 1$, a rede gerada pelo modelo WS tende a uma rede aleatória, mas não no sentido ER.	35
Figura 7 – (a) Rede gerada pelo modelo BA com $N = 300$ e $m = 3$. É possível observar a aparição de hubs no centro da rede. (b) Distribuição do grau para uma rede BA com $N = 10000$ e $m = 5$	36
Figura 8 – Resposta temporal dos modelos para propagação de epidemias.	38
Figura 9 – DAG representando uma cascata. Os vértices A e D são fontes. Os vértices B, C e H são <i>early adopters</i> , enquanto que os vértices E, F e I são folhas.	46
Figura 11 – (a) Rede de influência G . (b) Cascata inferida por V em G	49
Figura 12 – Distribuição de graus da rede de influências obtida. (a) Grau de Entrada (b) Grau de Saída	51
Figura 13 – Evolução temporal da rede. (a) <i>Clustering</i> . (b) Grau Médio. (c) <i>Mixing Assortativity</i>	51
Figura 14 – Distribuição de quantidade de usuários únicos por <i>hashtag</i>	52
Figura 15 – Popularidade acumulada das 6 <i>hashtags</i> mais compartilhadas no período observado.	53
Figura 16 – Resultado da aplicação da $SWT(80\%)$ a algumas <i>hashtags</i>	55
Figura 20 – Distribuição de HC para todas as <i>hashtags</i>	58
Figura 23 – Distribuição da diferença entre caminho mínimo e caminho obtido.	60
Figura 24 – <i>Betweenness centrality</i> como função do Grau.	61
Figura 25 – Pequena amostra da rede de co-ocorrências de <i>hashtags</i> . Cores correspondem à comunidades.	64
Figura 26 – Distribuição de D_g para várias amostras de G_r	66

LISTA DE ALGORITMOS

Algoritmo 1 – Algoritmo para calcular o número de caminhos disjuntos em um <i>DAG</i> . . .	47
Algoritmo 2 – Algoritmo para calcular a cascata <i>C</i> à partir de uma série de eventos <i>E</i> e uma rede de influências <i>G</i>	50
Algoritmo 3 – Algoritmo para calcular a menor janela de tempo que contém uma proporção <i>p</i> das observações	54

LISTA DE TABELAS

Tabela 1 – Principais características da rede de influências obtida.	50
Tabela 2 – Dez <i>hashtags</i> mais compartilhadas no período observado.	52
Tabela 3 – Quantidade de <i>hashtags</i> observadas e compartilhamentos, totais e para <i>hash-</i> <i>tags</i> com pelo menos 1000 compartilhamentos.	55
Tabela 4 – Dez <i>hashtags</i> mais compartilhadas no período observado e tamanho de cascata relacionado.	56
Tabela 5 – Exemplos de <i>hashtags</i> agrupadas na mesma comunidade.	65

SUMÁRIO

1	INTRODUÇÃO	23
1.1	Objetivos	24
1.2	Organização do Trabalho	25
2	CONCEITOS E MÉTODOS	27
2.1	<i>Dataset</i>	27
2.2	Redes Complexas	28
2.2.1	<i>Representação</i>	28
2.2.2	<i>Medidas de Caracterização</i>	29
2.2.3	<i>Modelos de Redes Aleatórias</i>	34
2.3	Processos dinâmicos em redes complexas	36
2.3.1	<i>Propagação de epidemias</i>	36
2.3.2	<i>Propagação de rumores</i>	38
3	TRABALHOS RELACIONADOS	41
3.1	Estudos no <i>Twitter</i>	41
3.2	Difusão de Informação em Redes	42
3.3	Homofilia	44
4	CASCATAS DE INFORMAÇÃO	45
4.1	Cascatas como Grafo Direcionados Acíclicos	45
4.2	Procedimentos e Dados Experimentais	49
4.2.1	<i>Rede de Influências</i>	49
4.2.2	<i>Difusão de Hashtags</i>	51
4.3	Caracterização de Cascatas	54
4.4	Caminhos Mínimos e Centralidade	59
4.5	Conclusão	62
5	HOMOFILIA E REDES DE TÓPICOS	63
5.1	Detecção de Tópicos	63
5.2	Homofilia e Topologia de Rede no <i>Twitter</i>	65
5.3	Conclusão	67
6	CONCLUSÃO E TRABALHOS FUTUROS	69

REFERÊNCIAS	71
-----------------------	----

INTRODUÇÃO

O começo do século XXI têm sido de mudanças profundas na forma como o ser humano se comunica e produz conhecimento, impulsionadas pela evolução da tecnologia de informação e a subsequente reorganização da sociedade em função dessa evolução. Economistas e sociólogos vêm estudando interações entre pessoas, mercados e instituições há mais de cinquenta anos (SIMON, 1955). Apesar disso, a natureza imprevisível dessas interações torna difícil uma observação prolongada desses sistemas, limitando esses estudos a pequenos grupos e períodos controlados. Um exemplo clássico de estudo com pequenos grupos fechados é o estudo por (ZACHARY, 1977), realizado com um pequeno grupo de karatê de até 100 membros durante um período de dois anos.

O aumento exponencial na quantidade de dados disponíveis sobre o comportamento e as interações humanas, provocado pelo advento e popularização da Internet, das mídias sociais, do comércio eletrônico e das plataformas móveis alterou essa situação, tornando possível a observação prolongada de mercados e redes sociais, por exemplo (WATTS, 2007). A esse aumento no volume, variedade e rapidez de produção de dados denominou-se *Big Data* (MCAFEE; BRYNJOLFSSON, 2012). A natureza do *Big Data* é heterogênea e multi-disciplinar por construção: muitos desses dados são produzidos diretamente por pessoas fazendo uso de algum serviço *online* (MANYIKA *et al.*, 2011). Dados como conversas em redes sociais, buscas em motores de pesquisa, conversas em sites de *micro-blogging* ou compras em lojas de *e-commerce* são gerados aos milhões a cada minuto. Um dos grandes desafios desse século é extrair informação e sentido desses dados (LYNCH, 2008).

Sistemas que são caracterizados por uma coleção de agentes que interagem entre si de alguma maneira, como é o caso de redes sociais, podem ser abordados como sistemas complexos (MITCHELL, 2006). Sistemas complexos são formados por partes discretas que interagem de forma não-linear, formando padrões de conexões não triviais cuja evolução ocorre sem um controle central (AMARAL; OTTINO, 2004). Características fundamentais desses sistemas

incluem emergência de comportamento coletivo, estrutura hierarquia e auto-organização (BARABÁSI, 2003). Pode-se citar como exemplos de sistemas complexos nosso sistema imunológico, o cérebro, a Internet, redes sociais, cadeias alimentares e o mercado financeiro (MITCHELL, 2009; COSTA *et al.*, 2011). Buscando melhor entender a estrutura - como as partes se conectam - e a dinâmica - como elas interagem - desses sistemas, foi desenvolvida a partir do final da década de 90 a Teoria das Redes Complexas (ALBERT; BARABÁSI, 2002; BARABÁSI; ALBERT, 1999). Por seu caráter extremamente multidisciplinar, os procedimentos de redes complexas vêm sendo utilizados nas mais diversas áreas do conhecimento, incluindo Física, Sociologia, Ciência da Computação, Biologia e Engenharia (COSTA *et al.*, 2011). Nesse trabalho, estamos interessados no estudo de redes de informação, tal como o *Twitter*.

Muitas das empresas líderes no mercado de tecnologia de informação fornecem parte de seus dados para pesquisa. O site de *micro-blogging Twitter*, por exemplo, possui uma *API* de *streaming* que fornece uma amostra aleatória de até 10% do conteúdo gerado na plataforma, que consiste de mensagens compartilhadas por usuários. Diversos estudos sobre interações humanas têm sido realizados com o auxílio desses dados, como em (BAKSHY *et al.*, 2011), que busca quantificar influência social no Twitter e (CATALDI; CARO; SCHIFANELLA, 2010), que propõe um método para detecção, em tempo real, de tópicos emergentes.

Nesse projeto, buscamos responder questões sobre a estrutura e a difusão de informação no *Twitter* e em redes sociais em geral. A natureza dos dados do *Twitter* proporciona diversos desafios, uma vez que a maioria dos algoritmos que desenvolvidos para análise de sistemas complexos possuem complexidade de tempo no mínimo quadrática em relação ao número de nós da rede, o que os torna impraticáveis para aplicação em *Big Data*. Com isso, nós propomos não apenas resolver questões diretamente relacionadas a esse conjunto de dados, mas também proporcionar uma coleção de métodos e conceitos que possam se demonstrar úteis em qualquer situação onde seja desejável o estudo de redes complexas derivadas de grandes conjuntos de dados reais.

1.1 Objetivos

Buscamos, com esse trabalho, responder as seguintes perguntas:

- Como podemos caracterizar o processo de difusão de informação no *Twitter* e numa rede social em geral? Que medidas são relevantes?
- Qual a influência da estrutura da rede nesse processo? Como distinguir um processo de difusão de informação que ocorre devido à rede de um que é apenas observado na rede?
- É possível agrupar tópicos utilizando apenas a topologia da rede? Até que ponto os tópicos compartilhados na rede influenciam sua topologia?

1.2 Organização do Trabalho

No capítulo 2 apresentamos o *dataset* utilizado, os métodos de obtenção de dados e conceitos fundamentais da teoria das redes complexas para esse trabalho. No capítulo 3 apresentamos uma revisão da literatura, focando principalmente em três tópicos: trabalhos que utilizam o *Twitter*, trabalhos sobre o processo de difusão de informação em redes e homofilia em redes sociais. No Capítulo 4, apresentamos métodos e medidas desenvolvidos para caracterizar o processo de difusão de informação em redes, juntamente com resultados obtidos. No Capítulo seguinte, estudamos a influência da homofilia no *Twitter*. Por fim, o Capítulo 6 apresenta as conclusões gerais desse trabalho.

CONCEITOS E MÉTODOS

2.1 *Dataset*

O *Twitter*¹ é um sistema de micro-blogging e rede social que possibilita a milhões de pessoas compartilhar e ler pequenos textos ("*tweets*") com até 140 caracteres. Um usuário no *Twitter* pode se inscrever ("seguir") outras pessoas, formando uma rede dirigida. Os *tweets* são exibidos para todos os seguidores de um usuário. Um usuário pode também replicar ("*retweet*") uma mensagem compartilhada por alguém que segue-o para seus seguidores, mantendo o crédito da mensagem ao autor original. É possível mencionar um outro usuário em um *tweet*, utilizando o símbolo "@" (e.g. @BarackObama).

Outro recurso bastante importante do *Twitter* são as *hashtags*. Usuários podem marcar suas postagens com identificadores de tópicos - as *hashtags* - textos ou frases que acompanhadas do símbolo "#" (e.g. #GoldenGlobes). Nesse trabalho, quando nos referindo ao *Twitter*, tratamos *hashtag* e tópico como sinônimos. Consideramos que um *retweet* que contém uma *hashtag* carrega um tópico de usuário para usuário, formando uma rede de difusão.

Utilizamos duas *APIS* para coletar dados do *Twitter*. A *API* de streaming² fornece gratuitamente uma amostra de aproximadamente 10% de *tweets* públicos em tempo real. A *API* Get Followers³ fornece dados sobre usuários, em especial quem são seus seguidores e quem ele segue. Os dados que analisamos foram coletados entre 22 e 26 de abril de 2016, correspondendo a um período de 5 dias.

Consideramos a estrutura de rede do *Twitter* de duas principais formas: (i) as relações entre usuários, juntamente com as informações de seguidores, *retweets* e citações e (ii) a relação entre *hashtags*. Com esses dados, estudamos a estrutura dessas redes e os processos dinâmicos

¹ <http://www.twitter.com>

² <https://dev.twitter.com/streaming/overview>

³ <https://dev.twitter.com/rest/reference/get/followers/ids>

Figura 1 – Possíveis formas de conexão no *Twitter*: (a) Dois usuários estão conectados se um segue o outro. (b) Dois usuários estão conectados se um menciona o outro. (c) Dois usuários estão conectados se um retweeta um post do outro. (d) Duas *hashtags* estão conectadas se compartilhadas no mesmo *tweet*. Nesse caso, o peso da conexão pode ser a quantidade de ocorrências. As relações em **negrito** podem ser derivadas diretamente do *tweet* de exemplo, onde Alice cita Bob e as *hashtags* café e cozinha no mesmo *post*. A relação de seguidor pode ser derivada, enquanto que o exemplo não possui nenhuma relação de *retweet*.

que ocorrem nelas, como o processo de difusão de um tópico. Exemplos de redes que podem ser formadas utilizando a estrutura do *Twitter* podem ser vistos na figura Figura 1.

Um exemplo do pacote de dados fornecido pela *API streaming* do *Twitter* pode ser visto na Figura 2. Os dados são enviados em um único pacote *json* contendo todas as informações disponíveis sobre o *tweet* (o usuário e o ID desse exemplo são fictícios).

O *Twitter* também fornece uma API que permite acesso ao conjunto total de *tweets* compartilhados, chamada *Firehose*, mas que não é disponibilizada para acesso livre. Um estudo quantitativo sobre a relevância estatística da API de *Streaming*, utilizando os dados obtidos com a *Firehose* em um mesmo período, foi realizado em (MORSTATTER *et al.*, 2013). Foi observado que a API de *Streaming* obtém bons resultados para aproximar informações como a distribuição de tópicos e medidas de topologia de rede, assumindo que os dados sejam coletados durante um período de mais de um dia.

2.2 Redes Complexas

2.2.1 Representação

A estrutura de sistemas complexos pode ser representada matematicamente por meio de grafos. Um grafo G é definido por um conjunto $V(G)$, de vértices (ou nós), um conjunto $E(G)$, arestas (ou *links*) onde cada elemento de E conecta dois elementos de V . Ocasionalmente, cada aresta pode possuir um peso associado, que é representado pelo mapeamento $W : E(G) \rightarrow \mathbb{R}$. Cada vértice de um grafo pode ser identificado por um número inteiro $i = 1, 2, \dots, N$, onde

```

1 {
2   'contributors ': None,
3   'truncated ': False,
4   'text ': 'Perdon por todo, nunca te quise lastimar #amor',
5   'in_reply_to_status_id ': None,
6   'id ': -,
7   'favorite_count ': 0,
8   'retweeted ': False,
9   'coordinates ': None,
10  'timestamp_ms ': u'1424721102661',
11  'entities ': {
12    'user_mentions ': [],
13    'symbols ': [],
14    'trends ': [],
15    'hashtags ': [amor],
16    'urls ': []
17  },
18  'in_reply_to_screen_name ': None,
19  'id_str ': -,
20  'retweet_count ': 0,
21  'in_reply_to_user_id ': None,
22  'favorited ': False,
23  'user ': {
24    'follow_request_sent ': None
25  }

```

Figura 2 – Exemplo de pacote *json* contendo um *tweet*.

$N = |V(G)|$. Cada aresta pode ser identificada por um par (i, j) , onde o primeiro elemento do par indica a origem da aresta e o segundo elemento indica o destino, se o grafo for dirigido. No caso não dirigido, a ordem no par (i, j) não importa.

Um grafo pode ser representado por sua matriz de adjacência. A matriz de adjacência A de um grafo G é uma matriz $N \times N$, onde a posição a_{ij} da matriz representa a existência de uma aresta do direcionada do vértice i ao vértice j , isto é, $a_{ij} = 1$ se i a j estão conectados ou $a_{ij} = 0$ caso contrário. Se um grafo é não dirigido, a existência da aresta (i, j) implica na existência da aresta (j, i) . A Figura 3 mostra um grafo não dirigido com sua matriz de adjacência correspondente.

2.2.2 Medidas de Caracterização

Redes complexas são, em geral, compostas por milhares, ou mesmo milhões, de vértices. Sendo assim, a inspeção visual não é suficiente para obter informações sobre sua organização, o que torna necessário utilizar medidas estruturais que auxiliem na caracterização, análise e classificação de redes complexas (BOCCALETTI *et al.*, 2006; NEWMAN, 2010). Estudos recentes têm

Figura 3 – Um grafo não-directionado e sua matriz de adjacência

se concentrado em uma pequena quantidade de propriedades que parecem ser comuns a muitas redes, cujo valor afeta seu funcionamento como sistema de forma fundamental (NEWMAN, 2003a).

O grau de um vértice, k_i , é dado pelo número de vértices a qual ele está conectado. Se considerarmos a representação matricial, o grau do vértice i é dado por $k_i = \sum_j a_{ij}$. Vértices com grau elevado são chamados de *hubs*. O grau médio $\langle k \rangle$ do grafo representa a densidade de conexões na rede.

$$\langle k \rangle = \frac{1}{N} \sum_i k_i. \quad (2.1)$$

Quando a rede é dirigida, o grau de um vértice pode ser separado em duas quantidades: o grau de entrada e o grau de saída. O grau de entrada é a contagem de quantas arestas chegam a um vértice, e é dado por

$$k_i^{in} = \frac{1}{N} \sum_j A_{ji}, \quad (2.2)$$

enquanto que o grau de saída é o número de arestas que saem do vértice

$$k_i^{out} = \frac{1}{N} \sum_j A_{ij}. \quad (2.3)$$

A organização das conexões de uma rede pode ser quantificada em termos de sua distribuição do número de conexões, $P(k)$, isto é, a probabilidade de que um vértice escolhido aleatoriamente tenha grau k . O nível de heterogeneidade (ou complexidade) na distribuição das conexões pode ser quantificado pela entropia de informação de Shannon, que é dada por:

$$H(k) = - \sum_{k_i=k_{min}}^{k_{max}} P(k_i) \log P(k_i), \quad (2.4)$$

Figura 4 – Ilustração esquemática de três situações onde o coeficiente de aglomeração tem diferentes valores. Considerando o vértice i em questão o vértice branco: Em (a) é apresentado um exemplo de clique, onde todos os vértices estão conectados entre si. Neste caso, $cc_i = 1$. Na figura (b), $cc_i = 3/10$. Já em (c) $cc_i = 0$, pois os vizinhos do vértice i não possuem conexões entre si.

onde o logaritmo é calculado na base 2. $H(k)$ é 0 quando todos os vértices têm o mesmo grau, como no caso de um grafo completo ou de um grafo regular (*grid*).

Em muitas redes reais, existe uma tendência a dois vértices que possuem conexão com um terceiro vértice possuírem também uma conexão entre si. Em redes sociais, isso implica dizer que duas pessoas com um amigo em comum têm mais chance de serem amigas do que duas pessoas escolhidas aleatoriamente (GIRVAN; NEWMAN, 2002).

O coeficiente de aglomeração local (ou *clustering*) traduz essa propriedade calculando o quão próxima está a vizinhança de um vértice de um clique (WATTS; STROGATZ, 1998). Seu valor é dado por

$$cc_i = \frac{2e_i}{k_i(k_i - 1)} = \frac{\sum_{j=1}^N \sum_{m=1}^N a_{ij}a_{jm}a_{mi}}{k_i(k_i - 1)}, \quad (2.5)$$

e_i representa o número de conexões entre os vizinhos do vértice i . Na figura 4 são apresentados três exemplos de configurações que geram coeficientes de aglomeração distintos.

O coeficiente de aglomeração traduz a tendência de formação de pequenos grupos densos em redes. Outra quantidade importante de ser estudada é a correlação entre propriedades de vértices adjacentes, o que é chamado de *assortative mixing*, ou assortatividade. Dizemos que uma rede é assortativa se há um viés para a formação de conexões entre vértices com características semelhantes, e disassortativa *disassortative mixing* se há a tendência de vértices a associarem-se com outros vértices de características diferentes. Laços de amizade costumam ser assortativos, já que são bastante afetados por raça, idade, posição social e língua, por exemplo (NEWMAN, 2003a). A assortatividade é também referida como *homofilia*.

Para calcular o *assortative mixing*, consideramos uma propriedade categórica D associada a cada vértice da rede. A quantidade d_{ij} representa a fração de arestas na rede que conecta um

vértice da categoria i a um vértice da categoria j . O coeficiente de assortatividade, r , é dado então por

$$r = \frac{\sum_i d_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}, \quad (2.6)$$

onde $a_i = \sum_j d_{ij}$ e $b_j = \sum_i d_{ij}$, as frações de todas as conexões que saem de vértices da categoria i e das conexões que chegam em vértices da categoria j , respectivamente. O valor desse coeficiente está entre $[-1, 1]$, onde -1 representa uma rede desassortativa e 1 uma rede assortativa. É comum usar E como o grau dos vértices.

Além da assortatividade, muitas redes apresentam estrutura de comunidades. Uma comunidade é um grupo de vértices com alta densidade interna de conexões, porém com uma densidade menor de conexões para fora do grupo (PALLA *et al.*, 2005). A noção de comunidades é intuitiva em redes sociais: pessoas se organizam em grupos naturalmente de acordo com seus interesses. Por exemplo, alunos frequentando a mesma escola, grupos de trabalho ou família costumam possuir muitas conexões dentro do mesmo ambiente. Muitas dessas comunidades apresentam estrutura hierárquica, com comunidades internas, como por exemplo alunos de uma mesma sala em uma escola ou times e funções diferentes dentro de uma empresa (NEWMAN; GIRVAN, 2004).

Apesar de intuitiva, existem diversas definições quantitativas para a noção de comunidades. A modularidade, Q , é uma quantidade associada a uma partição de uma rede em comunidades, e mede o quão divisiva é tal partição (GIRVAN; NEWMAN, 2002). Para isso, compara-se a proporção de vértices internos a uma partição com a proporção esperada em uma rede aleatória que segue o modelo de configuração

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{i,j} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \quad (2.7)$$

onde $a_{i,j}$ representa o peso da aresta entre os vértices i e j , $k_i = \sum_j A_{i,j}$ é a soma dos pesos de todos os vértices conectados ao vértice i , $m = \sum_{i,j} A_{i,j}$ é a soma de todos os pesos das arestas da rede, c_i e c_j são as comunidades do vértice i e j , e a função delta de Kronecker, isto é, $\delta(u, v) = 1$ se $u = v$ e 0 caso contrário.

Uma vez que apenas os únicos termos que afetam no cálculo da modularidade são os vértices de uma mesma comunidade, é possível reescrever a expressão 2.7 em termos de comunidades (FORTUNATO, 2010):

$$Q = \sum_{c=1}^{n_c} \left[\frac{l_c}{m} - \left(\frac{d_c}{2m} \right)^2 \right]. \quad (2.8)$$

Nesse caso, n_c é o número de comunidades, l_c é a soma dos pesos das arestas unindo vértices da comunidade c e d_c é a soma dos pesos de todas as arestas incidentes em c . A

modularidade retorna um valor entre 0 e 1, onde valores próximos de 1 representam uma partição melhor, ou seja, há estrutura de comunidades(FORTUNATO, 2010).

Menores caminhos, ou geodésicas, são um importante tópico de estudos em redes, em especial na análise de transportes e no processo de difusão de informação e comunicação(YAN *et al.*, 2006).

Um caminho entre i e j é uma sequência de vértices e arestas (nós adjacentes) que começa em i e termina em j , onde nenhum vértice é visitado mais de uma vez (BONDY; MURTY, 1976). Os menores caminhos entre todos os vértices em uma rede podem ser representados através de uma matriz de distâncias D , cujos elementos d_{ij} expressam o valor do comprimento do menor caminho entre os vértices i e j . Estes caminhos podem ser obtidos por meio dos algoritmos de Dijkstra ou uma busca em amplitude. O valor $d_{max} = \max_{i,j} d_{ij}$ é chamado diâmetro da rede. A média entre os valores na matriz D exprime o caminho característico da rede (comprimento médio do menor caminho), e é dada por

$$\ell = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}. \quad (2.9)$$

Caso i e j não pertençam a um mesmo componente (um componente é um conjunto de vértices tal que exista ao menos um caminho entre todos eles) conectado, $d_{ij} = \infty$. Logo, geralmente considera-se apenas o maior componente conectado na caracterização de uma rede.

Quando há transporte em uma rede, alguns vértices ou arestas recebem um tráfego mais intenso do que outros. Tais elementos representam os chamados “gargalos” e estão situados entre muitos dos menores caminhos, caso a informação trafegue por geodésicas. Quando removidos estes vértices, podem ocorrer rupturas na estrutura da rede, surgindo componentes não conectados, que são formados por vértices densamente conectados entre si, mas não conectados com o restante da rede. Para medir o tráfego que passa em um dado vértice (ou aresta), é usada a medida chamada *grau de intermediação* (*betweenness centrality*, em inglês) (NEWMAN, 2010), que mede o quanto um vértice ou aresta está no caminho mínimo entre outros vértices, e é calculada da seguinte forma

$$B_u = \sum_{ij} \frac{\sigma(i, u, j)}{\sigma(i, j)}, \quad (2.10)$$

$\sigma(i, u, j)$ é o número de menores caminhos entre os vértices i e j que passam pelo vértice (ou aresta) u e $\sigma(i, j)$ é o número total de menores caminhos entre i e j . A soma é feita sobre todos os pares distintos i, j de vértices. A média do grau de intermediação (*betweenness centrality*, em inglês) pode ser utilizada como uma medida de caracterização global da rede,

$$\langle B \rangle = \frac{1}{N} \sum_i B_i. \quad (2.11)$$

Figura 5 – (a) Rede aleatória com $N = 300$ e $p = 0.1$ (b) Distribuição do grau para uma rede ER com 10000 vértices e $p = 0.5$

O grau de intermediação assume que a informação trafega por menores caminhos. Outras medidas de centralidade importantes são as medidas *closeness centrality*, *PageRank*, *k-core* e *eigenvector centrality* (FREEMAN, 1978).

2.2.3 Modelos de Redes Aleatórias

A modelagem matemática de redes complexas teve início com Erdős e Rényi, com o intuito inicial de estudar, através de modelos probabilísticos, propriedades de grafos em função do número de conexões aleatórias presentes. Para isso, foi proposto um modelo de grafo aleatório de Erdős e Rényi (ER) (ERDÖS; RÉNYI, 1959). Nesse modelo, as arestas de um grafo G com N vértices são escolhidas aleatoriamente entre todas as $\frac{N(N-1)}{2}$ possíveis arestas. A probabilidade p de uma aresta ser escolhida é um parâmetro do modelo. A figura 5 mostra um exemplo desse tipo de rede.

A probabilidade do grau do i -ésimo vértice, k_i , possuir valor k nessa rede segue uma distribuição binomial com parâmetros $N - 1$ e p

$$2$$

O grau médio de uma rede aleatória será então a média de uma distribuição binomial, isto é, $\langle k \rangle = np$. Quando $N \rightarrow \infty$, a quantidade de vértices com grau k na rede segue uma distribuição de Poisson (ALBERT; BARABÁSI, 2002). A figura 5 mostra graus médios dos vértices em uma rede aleatória com $N = 10000$ e $p = 0.5$.

Embora muito estudado — principalmente por ser matematicamente tratável — o modelo ER não é o mais adequado para reproduzir a maior parte das redes reais. Uma propriedade que muitas redes reais exibem, como as redes sociais, é a propriedade do *pequeno mundo* - a maior parte dos vértices pode ser atingido a partir de todos os outros passando por um pequeno número

de arestas (MILGRAM, 1967). Redes ER apresentam a propriedade de pequeno mundo, porém seu coeficiente de agrupamento é muito baixo quando comparado a redes reais (NEWMAN, 2003b).

Watts e Strogatz (WATTS; STROGATZ, 1998) propuseram um modelo de rede aleatória chamado *small-world* de Watts-Strogatz (WS) que apresenta um coeficiente de agrupamento (Equação 2.5) alto. O modelo aceita três parâmetros: N , a quantidade de vértices da rede; K , a quantidade de ligações iniciais de um vértice e p , a probabilidade de mudar uma aresta do grafo. A construção da rede parte de um anel, onde cada vértice se conecta inicialmente aos $\frac{K}{2}$ vizinhos em cada um de seus lados. Após esse passo inicial, cada aresta é modificada com uma probabilidade p , diminuindo as distâncias na rede. A Figura 6 mostra redes geradas para $N = 25$, $K = 4$ para diferentes valores de p . A distribuição do grau numa rede WS é similar a ER. A Figura 6 mostra a evolução de uma rede em anel para uma rede aleatória ao se variar o parâmetro p do modelo WS.

Figura 6 – Redes geradas pelo modelo WS com $K = 4$ e $N = 25$ quando (a) $p = 0.0$, (b) $p = 0.2$, (c) $p = 0.5$ e (d) $p = 0.8$. Quando $p \rightarrow 1$, a rede gerada pelo modelo WS tende a uma rede aleatória, mas não no sentido ER.

Por fim, Barabási e Albert (BARABÁSI; ALBERT, 1999) mostraram que a distribuição do grau em uma rede real é muitas vezes assimétrica, com alguns vértices (*hubs*) com muitas conexões. Eles propuseram o modelo BA, no qual uma rede inicial com N_0 vértices, todos conectados entre si, aumenta a cada passo. Cada novo vértice adicionado conecta-se a m outros vértices. A probabilidade de um novo vértice i conectar-se a um vértice j já presente na rede é diretamente proporcional ao grau desse vértice, isto é,

$$P(i \rightarrow j) = \frac{k_j}{\sum_l k_l}. \quad (2.12)$$

A Figura 7 mostra uma rede BA com $N = 300$ vértices e a distribuição do grau dos vértices para uma rede BA com $N = 10000$ graus e $m = 5$. Devido à ligação preferencial, o modelo produz a lei de potência na distribuição dos graus dos vértices observada na maior parte das redes reais.

Figura 7 – (a) Rede gerada pelo modelo BA com $N = 300$ e $m = 3$. É possível observar a aparição de hubs no centro da rede. (b) Distribuição do grau para uma rede BA com $N = 10000$ e $m = 5$

2.3 Processos dinâmicos em redes complexas

Para se entender um sistema, é necessário conhecer a sua organização. Definidas as medidas de redes que utilizamos em nosso trabalho, descrevemos a seguir processos dinâmicos que tomamos como base para analisar as redes sociais. Apesar de não utilizarmos propagação de epidemias diretamente, a definimos aqui pois sua base teórica é importante para outros processos dinâmicos de propagação de informação que iremos considerar.

2.3.1 Propagação de epidemias

Suponha que uma pessoa infectada transmite uma determinada doença com probabilidade $\beta \langle k \rangle$, significando que cada indivíduo tem, na média, $\langle k \rangle$ contatos com outros indivíduos por unidade de tempo. Definindo S e I como o número médio de indivíduos susceptíveis e infectados, respectivamente, e considerando uma população de N indivíduos, define-se $s = S/N$ como a fração indivíduos susceptíveis. Uma vez que há na média uma fração de $i = \frac{I}{N}$ indivíduos infectados no total, isto implica que a média global da taxa de novas infecções será $\beta \langle k \rangle si$. Assim, é possível expressar este modelo pela seguinte equação diferencial (BARRAT; BARTHLEMY; VESPIGNANI, 2008),

$$\frac{di(t)}{dt} = \beta \langle k \rangle i(t) [1 - i(t)], \quad (2.13)$$

onde se utilizou a relação $s + i = 1$. A solução desta equação pode ser obtida utilizando-se uma aproximação linear, isto é, negligenciando-se os termos $O(i^2)$. Sua solução exata é dada por (BARRAT; BARTHLEMY; VESPIGNANI, 2008),

$$i(t) = \frac{x_0 e^{t/\tau}}{1 + i_0(e^{t/\tau} - 1)}, \quad (2.14)$$

onde i_0 é a densidade inicial de indivíduos infectados e $\tau = (\beta \langle k \rangle)^{-1}$.

O modelo SIS é guiado por uma equação semelhante a equação 2.13, incluindo apenas um termo de transição espontânea com taxa μ . Assim, a dinâmica do modelo é governada por por (BARRAT; BARTHLEMY; VESPIGNANI, 2008),

$$\frac{di(t)}{dt} = -\mu i(t) + \beta \langle k \rangle i(t)[1 - i(t)]. \quad (2.15)$$

É interessante notar que o termo $-\mu i(t)$ não depende do grau médio da rede, pois trata-se de um processo espontâneo, onde o indivíduo passa de infectado à susceptível novamente. O modelo SIR (ANDERSON; MAY, 1992) é semelhante, porém, após este processo de recuperação o indivíduo passa à um terceiro estado, no qual ele está recuperado e não participa de qualquer interação. Esta dinâmica é regida pelo seguinte sistema de equações diferenciais (BARRAT; BARTHLEMY; VESPIGNANI, 2008),

$$\begin{cases} \frac{ds(t)}{dt} = -\beta \langle k \rangle i(t)[1 - r(t) - i(t)], \\ \frac{di(t)}{dt} = -\mu i(t) + \beta \langle k \rangle i(t)[1 - r(t) - i(t)], \\ \frac{dr(t)}{dt} = \mu i(t), \end{cases} \quad (2.16)$$

onde $r(t)$ são os indivíduos recuperados e esta recuperação ocorre com uma taxa μ .

A Figura 8 exemplifica o comportamento destas três variáveis (S, I, R) considerando uma rede totalmente conectada para os modelos SI, SIR e SIS.

Tanto o modelo SIS quanto o modelo SIR apresentam comportamento semelhantes para a fração de indivíduos infectados para valores pequenos de t . Isto ocorre, pois a equação de variação do número de infectados $\frac{di(t)}{dt}$ é a mesma para ambas dinâmicas (BARRAT; BARTHLEMY; VESPIGNANI, 2008). Utilizando uma metodologia semelhante a realizada para a dinâmica SI com uma aproximação linear, negligenciando os termos $O(i^2)$ para valores pequenos de t , têm-se,

$$\frac{di(t)}{dt} = -\mu i(t) + \beta \langle k \rangle i(t), \quad (2.17)$$

para o modelo SIR deve-se considerar que o termo $r(t)$ pode ser considerado de mesma ordem de grandeza de $i(t)$ para valores próximos de $t = 0$. A solução desta é da forma $i(t) \simeq i_0 e^{t/\tau}$, onde i_0 é a densidade inicial de indivíduos infectados e

$$\tau^{-1} = \beta \langle k \rangle - \mu. \quad (2.18)$$

Esta relação fornece o chamado limiar epidêmico, $\frac{\mu}{\beta} > \langle k \rangle$ (pois $\tau > 0$). Assim, caso esta relação não seja satisfeita, a doença tende a ser eliminada em um tempo finito.

As equações apresentadas até aqui consideraram apenas redes regulares ou onde o grau decai exponencialmente ao distanciar-se do grau médio, $\langle k \rangle$, que ocorre em redes de modelos aleatório de Erdős-Renyi e de pequeno mundo de Watts-Strogatz. Além dos resultados apresentados aqui é possível avaliar estas dinâmicas em redes heterogêneas e com correlação de

Figura 8 – Resposta temporal dos modelos para propagação de epidemias. (a) Modelo SI: um pequeno número de indivíduos infectados inicialmente (1% neste exemplo), apresentará um crescimento exponencial no início, chegando a saturação em 1 no estado estacionário; (b) Modelo SIR: mostra as frações da população suscetíveis, infectadas e recuperadas. Os parâmetros desta simulação foram $\beta = 1$, $\gamma = 0.2$, $s_0 = 0.99$, $x_0 = 0.01$ e $r_0 = 0$; (c) Modelo SIS: a fração dos indivíduos infectados neste modelo cresce de acordo com a curva logística, de maneira semelhante ao modelo SI, entretanto, a fração de infectados nunca chega a 1. Figuras adaptadas de (NEWMAN, 2010)

grau, isto pode ser encontrado na referência (BARRAT; BARTHLEMY; VESPIGNANI, 2008). Entretanto, considera-se apenas estatísticas de primeira ordem, desconsiderando a estrutura global da rede, como comunidades por exemplo.

2.3.2 Propagação de rumores

Outra dinâmica de grande interesse da comunidade científica é a propagação de rumores. Esta dinâmica é semelhante a descrita na seção anterior, 2.3.1, porém neste caso o interesse é propagar a informação da maneira mais eficiente possível. O modelo padrão para propagação de rumor foi proposto há algumas décadas por Daley e Kendall (DALEY; KENDALL, 1964), chamado modelo DK. Recentemente, muitos autores têm explorado este tópico com a aplicação

de topologias de redes complexas. Este modelo possui três estados, denotados por I (ignorante), S (informantes, do inglês *spreader*) e R (contidos, do inglês *stifler*). Os ignorantes são aqueles indivíduos que não conhecem o rumor e são susceptíveis à informação. Os espalhadores são aqueles que já ouviram o rumor e estão espalhando o mesmo. Finalmente, contidos são aqueles que conhecem o rumor, mas não o espalham mais, ou seja, já perderam interesse por tal rumor.

O processo de propagação evolui pelo contato entre um informantes e ignorantes. Quando um ignorante encontra um informante ele passa a ser um informante a uma taxa λ . O decaimento deste ocorre por meio pelo esquecimento ou pela perda do interesse no rumor, assim os informantes tornam-se contidos com uma probabilidade α se estiverem em contato com outros informantes ou contidos.

Esta dinâmica é semelhante à propagação de epidemias SIR, porém, neste caso a recuperação é substituída pela perda de interesse no rumor e isto não ocorre de maneira espontânea, mas sim pelo contato entre informante e outros informantes ou informantes e indivíduos contidos. Para uma distribuição homogênea, isto é onde o grau decai exponencialmente ao distanciar-se do valor médio, o conjunto de equações diferenciais que descrevem a evolução temporal desta é dado por (MORENO; NEKOVEE; PACHECO, 2004; NEKOVEE *et al.*, 2007; BARRAT; BARTHLEMY; VESPIGNANI, 2008):

$$\begin{cases} \frac{di(t)}{dt} = -\lambda \langle k \rangle i(t)s(t), \\ \frac{ds(t)}{dt} = \lambda \langle k \rangle i(t)s(t) - \alpha \langle k \rangle s(t)[s(t) + r(t)], \\ \frac{dr(t)}{dt} = \alpha \langle k \rangle s(t)[s(t) + r(t)], \end{cases} \quad (2.19)$$

onde λ é a taxa de propagação do rumor e α é a taxa de contenção do mesmo. Com algumas manipulações, é possível derivar o limiar de propagação, um conceito semelhante ao limiar epidêmico, porém aplicado ao contexto de propagação de rumores. Este é dado por,

$$\frac{\lambda}{\alpha} > 1. \quad (2.20)$$

Além desta análise em redes homogêneas, estas equações podem ser generalizadas para redes heterogêneas, de maneira semelhante a feita para as dinâmicas de epidemias na seção anteriores. Tal análise foi realizada nas referências (NEKOVEE *et al.*, 2007; BARRAT; BARTHLEMY; VESPIGNANI, 2008). Em (NEKOVEE *et al.*, 2007) é feita a análise destas equações derivando-se o limiar de propagação, considerando também um termo de esquecimento, no qual o indivíduo perde o interesse pelo rumor e passa de informante a contido espontaneamente.

Além do modelo de Daley e Kendall, outro modelo de grande interesse é o modelo de Maki-Thompson. Neste caso o contato é direcionado, ou seja, apenas o nó que inicia o contato pode sofrer alterações em seu estado, diferentemente do modelo DK, onde ambos podem mudar de estado. É importante enfatizar que estes dois modelos tendem a uma mesma resposta considerando a abordagem de campo médio, entretanto o modelo DK apresenta uma variância maior que o modelo MT (LEBENSZTAYN; MACHADO; RODRÍGUEZ, 2011).

TRABALHOS RELACIONADOS

Nesse capítulo, são apresentados alguns dos principais trabalhos relacionados ao *Twitter* à propagação de informação em redes.

3.1 Estudos no *Twitter*

A plataforma de *micro-blogging*, *Twitter*, vêm sendo estudada em trabalhos científicos exaustivamente desde sua popularização, devido principalmente a disponibilidade de dados de uso da rede pela própria empresa.

O trabalho de (KWAK *et al.*, 2010) foi um dos primeiros a abordar questões quantitativas sobre a plataforma. O trabalho mostra que a maioria ($> 85\%$) dos *trending topics* são relacionados a notícias - o que coloca a plataforma numa posição mais próxima a uma mídia de notícias do que uma rede social.

A identificação de vértices influentes é um problema importante em qualquer rede social, com aplicações em anúncios e *marketing viral*, por exemplo. No *Twitter*, a medida mais direta de influência é o número de seguidores de um usuário. Utilizar as medidas *Page Rank* e número de seguidores como medida de influência produz resultados muito parecidos, enquanto que o número de *retweets* difere das duas medidas anteriores (KWAK *et al.*, 2010). Em (CHA *et al.*, 2010), três medias de influência são analisadas: número de seguidores, *retweets* e menções. Ambos os trabalhos concluem que uma quantidade alta de seguidores não necessariamente implica numa maior capacidade de propagar informação. Além disso, usuários que falam bastante sobre um mesmo tópico possuem maior capacidade de propagação, em termos de *retweets* e menções obtidos.

Numa rede social *online*, como o *Twitter* ou *Facebook*, não é raro um usuário possuir muitos amigos ou seguir muitas pessoas. Isso não significa que aconteçam interações reais através dessas arestas. Huberman, Romero e Wu (2008) define uma relação de amizade no

Twitter como uma relação onde há ao menos uma comunicação direta entre dois usuários. Essa relação é melhor para prever atividade de um usuário na rede, em comparação com o número declarado de seguidores.

Como plataforma de mídia, os tópicos que são compartilhados no *Twitter* têm tanta, ou mais importância, que a estrutura de conexões da rede social que ele representa. Utilizando o *Page Rank* e as redes de seguidores e *retweets*, Welch *et al.* (2011) mostra que, *retweetar* um usuário é um indicador significativamente mais forte de interesse topical do que a relação de seguidor.

O trabalho de Zhao *et al.* (2011) compara os tópicos compartilhados no *Twitter* com tópicos do jornal *New York Times*, utilizando o método *latent dirichlet allocation (LDA)* para identificação de tópicos nos textos (BLEI; NG; JORDAN, 2003). O *LDA* também é empregado para classificar em tópicos o conteúdo obtido à partir de um recurso do *Twitter* chamado *user lists* - listas de interesse geralmente organizadas por *tags* (KANG; LERMAN, 2012). Outra forma de classificar tópicos no *Twitter* é utilizando diretamente as *hashtags* e a rede de co-ocorrências subjacente, como feito em (WENG, 2014). Esse último método é empregado nesse trabalho.

Um modelo para prever a capacidade de *retweets* de um *tweet* foi proposto por (SUH *et al.*, 2010). Foi constatado que as características mais importantes para prever a popularidade de um *tweet* são: *hashtags*, presença de *URLs* e o número de seguidores. Apesar disso, uma observação importante é a de que o número de *retweets* passados (histórico de usuário) têm pouco influência nesse processo. Zaman *et al.* (2010) propõe um modelo probabilístico para resolver o mesmo problema, e observa que a característica mais importante é a identidade dos usuários envolvidos. Esse modelo é treinado durante uma hora e é utilizado para prever a quantidade de *retweets* na hora seguinte.

3.2 Difusão de Informação em Redes

Processos dinâmicos em redes complexas, como os abordados na seção 2.3, vêm sendo estudados exaustivamente nas últimas décadas. Nesse trabalho, estamos interessados principalmente em processos de difusão de informação em redes.

Uma das primeiras idéias propostas para modelar o comportamento coletivo em difusão de informação é um modelo de decisões binárias com *thresholds*, ou limiares (GRANOVETTER, 1978). Nesse modelo, cada vértice pode ou não fazer parte do processo de propagação de informação. A decisão de participar ou não de um processo, para cada vértice, depende um *threshold* individual, θ_i , que representa a fração de seus vizinhos do i -simo vértice que precisa aderir ao processo para que ele também o faça. A simulação é feita considerando um conjunto de vértices iniciais V_i , que possuem a informação. O sistema evolui em intervalos de tempo discretos, testando a cada instante fração de vizinhos de cada vértice em relação ao *threshold*.

O modelo de decisão binária com limiares pode ser aplicado também a falhas em redes de transmissão, internet, percolação e voto por maioria, e difere dos modelos de propagação de epidemia tradicionais principalmente pois apresenta dependência local: a influência de um vértice A sobre outro vértice B depende do estado dos outros vizinhos de B . Uma solução exata para o modelo em grafos aleatórios é apresentada em (WATTS, 2002), e uma condição inicial necessária para a existência de cascatas globais identificada. Um método aproximado baseado em *hill climbing* para a escolha de vértices iniciais para o processo de difusão de informação que maximizam o tamanho final da cascata é apresentado em (KEMPE; KLEINBERG; TARDOS, 2003).

Um dos objetivos principais, com aplicações especiais em *marketing* viral, do estudo de propagação de informação é entender os efeitos da troca de informação *boca a boca*, onde a informação é passada de pessoa para pessoa ao longo da rede social. Um estudo sobre o *Flickr* realizado em 2009 sobre a rapidez com a qual imagens são disseminadas nessa rede social (CHA; MISLOVE; GUMMADI, 2009) mostrou que mesmo as imagens mais populares atingem usuários com distância maior que dois com relativa raridade (20%). O trabalho também apresenta um estudo sobre a evolução temporal na popularidade de imagens, que demonstram padrões parecidos aos observados na popularidade de *hashtags* no *Twitter*.

A informação obtida sobre difusão de informação em redes sociais *online* nem sempre é completa: muitas vezes, é necessário trabalhar com amostras parciais dos dados. Ao inferir uma cascata de informação, informações faltantes podem afetar inferências sobre o processo de difusão. Um método para prever características como tamanho e profundidade de uma cascata completa a partir de uma fração proveniente de uma amostra é proposto em (SADIKOV *et al.*, 2011). O método foi aplicado em uma rede obtida a partir do *Twitter* com 70 milhões de vértices. Em (RODRIGUEZ; LESKOVEC; KRAUSE, 2010) é proposto um método de otimização para inferir a topologia mais provável de uma rede não observável a partir de uma sequência de observações provenientes de vários processos de difusão de informação (vértices compartilhando alguma informação, como um usuário do *Twitter* ao escrever um *post* novo, por exemplo). Esse algoritmo utiliza a noção de que, em um processo de difusão, se um vértice v passa a possuir a informação logo após o vértice u , a chance de que a aresta $u \rightarrow v$ exista aumenta proporcionalmente à diferença de tempo observada.

A tarefa de prever características de uma cascata é bastante complicada. Em geral, cascatas com um grande número de vértices são relativamente raras, enquanto que a maioria dos processos de difusão de informação forma pequenas árvores, onde todos os vértices estão a um ou dois passos da origem da informação (GOEL; WATTS; GOLDSTEIN, 2012). Ainda é um problema em aberto determinar se um processo de difusão observado em uma rede depende de conexões internas ou de agentes externos, como mídias de massa.

3.3 Homofilia

O termo homofilia¹ (do inglês *homophily*) refere-se a tendência de indivíduos a associar-se com indivíduos parecidos em respeito a alguma relação de similaridade: idade, interesses, religião e etnia, por exemplo. O conceito foi inicialmente proposto em (MCPHERSON; SMITH-LOVIN; COOK, 2001).

Diversos estudos empíricos vêm sendo realizados sobre homofilia em redes sociais, em diferentes domínios, como em *relacionamentos online* (FIORE; DONATH, 2005), *Flickr* e *Last.fm* (AIELLO *et al.*, 2012) e *Facebook* (WENG; LENTO, 2014). Esses estudos têm como objetivo identificar e quantificar a presença de homofilia nessas plataformas. Por exemplo, Weng e Lento (2014) mostra que, no *Facebook*, usuários com interesses similares formam sub-redes densamente conectadas identificadas por tópicos, com pouca sobreposição.

Rogers (2010) sugere que ao menos algum grau de *heterofilia* (diferenças em crenças, posição social, *hobbies*, etc) deve existir entre os envolvidos em processos de difusão de informação *nova* (inovações), enquanto que a maior parte do restante das comunicações ocorrem em entre indivíduos que possuem interesse comum. A distinção entre influência e homofilia muitas vezes é bastante complicada. Pessoas com interesses similares tendem a se relacionar, e, ao mesmo tempo, grupos sociais tendem a se tornar mais similares devido ao efeito da influência social (*peer influence*) (SHALIZI; THOMAS, 2011). Esse efeito é explicado e estudado em (ARAL; MUCHNIK; SUNDARARAJAN, 2009), que mostra que, em uma rede de mensagens instantâneas, a homofilia é responsável por promover a maior parte dos processos de difusão de informação, ao invés da influência direta.

¹ Utilizamos o termo "homofilia" em português nesse trabalho com um sentido diferente do sentido apresentado na maioria dos dicionários.

CASCATAS DE INFORMAÇÃO

Nesse capítulo, buscamos responder as perguntas apresentadas na introdução:

- Como podemos caracterizar o processo de difusão de informação no *Twitter* e numa rede social em geral? Que medidas são relevantes?
- Qual a influência da estrutura da rede nesse processo? Como distinguir um processo de difusão de informação que ocorre devido à rede de um que é apenas observado na rede?

Na seção 4.1 apresentamos algumas medidas desenvolvidas para caracterizar cascatas de informação. Na seção 4.2 apresentamos e caracterizamos os dados utilizados nos experimentos subsequentes.

4.1 Cascatas como Grafo Direcionados Acíclicos

Enquanto que os padrões de conexões em redes reais podem ser representados geralmente por grafos gerais, processos de propagação de informação em redes formam cascatas, que podem ser representadas como grafos direcionados acíclicos (*DAG - directed acyclic graphs*). Nessas cascatas, vértices representam agentes e arestas direcionadas representam um caminho tomado por uma informação, que pode ser uma notícia, um meme ou um vírus, por exemplo ([SADIKOV et al., 2011](#)).

Algumas propriedades dos *DAGs* podem ser úteis para caracterizar cascatas de informação. Consideremos o *DAG* da Figura 9, onde vértices são usuários do *Twitter* e arestas representam propagação de uma notícia. Chamamos de *fontes*, ou *innovators* todos os vértices que possuem grau de entrada 0 — isso significa que esse vértice não recebeu a notícia de nenhuma outra fonte interna à rede. Todo vértice que recebe uma informação direto da fonte é chamado de *early adopter* ([WATTS, 2002](#)). Chamamos de *folhas* vértices que possuem grau

Figura 9 – DAG representando uma cascata. Os vértices A e D são fontes. Os vértices B, C e H são *early adopters*, enquanto que os vértices E, F e I são folhas.

de saída 0, isto é, vértices onde a propagação da informação termina. Todos os outros vértices intermediários são parte do processo de propagação.

O *alcance*, A , de uma cascata é definido como o caminho mais longo entre um vértice fonte e uma folha. Podemos calcular o alcance parcial de um vértice v utilizando a seguinte recorrência

$$A_v = \operatorname{argmax}_{x \in V(v)} A_x \quad (4.1)$$

$V(v)$ representa o conjunto de vizinhos do vértice v . Para calcular A basta tomar o maior A_x do grafo. O alcance é importante em cascatas de informação, pois mede o quão longe uma informação consegue chegar a partir da fonte. Se considerarmos a propriedade de *pequeno mundo* (MILGRAM, 1967), esperamos que o valor de A seja baixo para redes sociais ($A \leq 6$). Um estudo foi feito com o *Facebook*, em 2012, e a distância média encontrada entre usuários foi aproximadamente igual a quatro (BACKSTROM *et al.*, 2012).

Uma cascata de informação geralmente trata de apenas um tópico. Entretanto, ela pode surgir de diferentes fontes. Definimos como índice de heterogeneidade, HC , de uma cascata como o número máximo de caminhos disjuntos (que não compartilham vértices) entre os vértices fonte e as folhas. Se considerarmos que uma informação que trafega pela rede é uma função dos vértices (ou pessoas) pelos quais ela passou, o número de caminhos disjuntos é o equivalente a medir quantas versões completamente diferentes existem de um assunto. No caso da Figura 9, $HC = 2$, considerando os caminhos $(A \rightarrow B \rightarrow E)$ e $(D \rightarrow H \rightarrow I)$, por exemplo. Isso significa que, mesmo que I e E estejam falando sobre o mesmo tópico, suas versões podem ser totalmente

diferentes.

Para calcular o número de caminhos disjuntos no *DAG*, utilizamos o *fluxo máximo* e o algoritmo de *Edmonds e Karp* (EDMONDS; KARP, 1972). O método está descrito no Algoritmo 1. Todas as capacidades das arestas são unitárias.

Algoritmo 1: Algoritmo para calcular o número de caminhos disjuntos em um *DAG*.

Input: Cascata C em forma de *DAG*
Output: $HC(C)$

```

1  $g \leftarrow$  grafo vazio
2 cria vertice ( $g$ ,  $source$ )
3 cria vertice ( $g$ ,  $sink$ )
4 for  $v \in V(C)$  do
5   | cria vertice ( $g$ ,  $v_{in}$ )
6   | cria vertice ( $g$ ,  $v_{out}$ )
7   | cria aresta ( $g$ ,  $v_{in} \rightarrow v_{out}$ )
8   | if  $outdegree(v) = 0$  then
9   |   | cria aresta ( $g$ ,  $v_{out} \rightarrow sink$ )
10  | if  $indegree(v) = 0$  then
11  |   | cria aresta ( $g$ ,  $source \rightarrow v_{in}$ )
12 for  $e \in E(C)$  do
13   | Seja  $e : A \rightarrow B$ 
14   | cria aresta ( $g$ ,  $A_{out} \rightarrow B_{in}$ )
15 return  $maxflow(g)$ 
```

O valor máximo para HC é o mínimo entre o número de fontes e o número de folhas. Note que ignoramos os vértices isolados no processo, pois não consideramos um caminho de tamanho 0. A complexidade desse algoritmo é a complexidade do algoritmo utilizado para o cálculo do fluxo máximo, que, no caso do algoritmo de *Edmonds e Karp*, é de $O(VE^2)$.

Uma rede em estrela, como a da Figura 10 (a), com um *hub* central e 5 vértices periféricos terá $HC = 1$, pois só há uma história em potencial, enquanto que na Figura 10 (b), apesar de existirem duas versões originais na história, também tem $HC = 1$, uma vez que a informação necessariamente passa por D .

Muitas vezes, o processo direto de difusão de informação não é inteiramente observável. Ao invés disso, conhecemos uma rede direcionada, G , onde cada aresta entre os vértices a e b indica que há uma relação de influência entre $a \rightarrow b$, ou seja, há um caminho para a informação se propagar. Além disso, conhecemos uma série ordenada de eventos, ε , onde cada evento representa um vértice de $v \in G$ e um tempo t , indicando que o vértice v participou do processo de difusão no tempo t . Algumas vezes o conjunto ε não é completo, de forma que a cascata inferida possui informação faltante.

O processo utilizado por nós para inferir a cascata C à partir de G e ε é o seguinte. Consideramos um evento $e_i \in \varepsilon$. Olhamos para todos os vértices na rede original G que possuem

Figura 10 – Exemplos de duas redes diferentes com $HC = 1$.

uma aresta para o vértice v_i relativo ao i -ésimo evento e que já apareceram em algum evento e_j com $j < i$. Consideramos que todos os vértices nessa condição *influenciaram* o vértice v_i a fazer parte do processo, de modo que a aresta passa a existir na cascata.

Por exemplo, considere G a rede da Figura 11 e os eventos $\varepsilon = \{C, A, H, F, I, L\}$. Note que G pode conter ciclos. A cascata inferida pode ser vista na Figura 11 (b). É possível notar que existe a possibilidade do vértice H ter sido influenciado indiretamente pelo vértice A , porém essa informação não está presente em ε , de forma que é impossível afirmar com certeza.

Algumas vezes, faz sentido considerar que existe a relação $A \rightarrow D \rightarrow H$, por exemplo, mesmo que D não tenha sido observado no processo de difusão. Vamos definir como vértice *intermediário* um vértice v para o qual existem outros dois vértices, a e b tais que existem as arestas $a \rightarrow v$ e $v \rightarrow b$, e a e b foram observados em ε . Chamaremos de P_i a probabilidade de um vértice intermediário fazer parte do processo de difusão. Desejamos obter a cascata C mais provável considerando essa probabilidade e a possibilidade de incluir vértices intermediários.

Vamos chamar de P_c a chance da cascata C inferida sobre G estar perfeitamente correta, ou seja, todos os vértices em C realmente fizeram parte do processo de difusão. Note que os vértices já observados não fazem parte do cálculo. Podemos notar que

$$P_c = \prod_{v \in V_i} P_i I(v \in C) + (1 - P_i) I(v \notin C), \quad (4.2)$$

onde V_i é o conjunto de vértices intermediários de C em relação a G e $I(e) = 1$ se o evento e ocorre. Vamos assumir que P_i é o mesmo para todos os vértices do processo de difusão. Dessa forma, a função P_c nada mais que um produtório de vários termos iguais no formato $P_i I(v \in C) + (1 - P_i) I(v \notin C)$. Para maximizar P_c , devemos maximizar cada termo individualmente - ou

Figura 11 – (a) Rede de influência G . (b) Cascata inferida por V em G .

seja, devemos incluir o vértice intermediário se $P_i > 0.5$ e não incluir caso contrário. Isso nos permite desenvolver o Algoritmo 2 para construir a cascata C a partir da rede G e dos eventos ε . No algoritmo, $Viz_{G_T}(v)$ representa a vizinhança do vértice v em G transposto. A complexidade do Algoritmo 2 é $O(|E||V(G)|^2)$, no pior caso, porém o algoritmo é mais rápido para redes esparsas. Em aplicações práticas, onde a probabilidade de que um vértice faça parte de uma cascata já é muito baixa, quase sempre a melhor opção é não adicionar vértices extras à cascata ($P_i < 0.5$). Detalhes de como estimar P_i são apresentados mais adiante.

4.2 Procedimentos e Dados Experimentais

4.2.1 Rede de Influências

Para construir nosso conjunto de dados, utilizamos a *API* da seguinte forma. Cada pacote fornecido pela *API* corresponde a um *tweet*, que contém, dentre outras, informações sobre outros usuários mencionados no *post*, data e *hashtags* compartilhadas (ver a Figura 2, que mostra o formato de um pacote).

Utilizamos a informação de menções a outros usuários para construir uma rede de influência G à medida que os dados foram sendo coletados. Quando um usuário a cita um usuário b em seu *tweet*, adicionamos a aresta $b \rightarrow a$ em G , representando que b exerce uma influência sobre a . A Tabela 1 mostra algumas medidas do estado de G após a coleta de dados. O grau

Algoritmo 2: Algoritmo para calcular a cascata C à partir de uma série de eventos E e uma rede de influências G .

Input: Rede de influências G .

Input: Série de eventos E ordenada temporalmente.

Input: Probabilidade Q de um vértice intermediário participar do processo.

Output: Cascata C .

```

1  $K \leftarrow \{\}$ 
2  $C \leftarrow$  grafo vazio
3 for  $e \in E$  do
4   if  $v \notin K$  then
5      $v \leftarrow e.\text{vertice}$ 
6     for  $x \in \text{Viz}_{G_T}(v)$  do
7       if  $x \in K$  then
8          $C \leftarrow \text{aresta}\{x \rightarrow v\}$ 
9       else
10        if  $P_i > 0,5$  then
11          for  $y \in \text{Viz}_{G_T}(x)$  do
12            if  $y \in K$  then
13               $C \leftarrow \text{aresta}\{y \rightarrow x\}$ 
14               $C \leftarrow \text{aresta}\{x \rightarrow v\}$ 
15    $K \leftarrow K + \{v\}$ 
16 return  $C$ 

```

Medida	#Vértices	#Arestas	Assortativity Coefficient	$\langle k \rangle$	Entropia	Clustering Global
Valor	7216591	9591104	-0.007	1.653	0.014	0.02

Tabela 1 – Principais características da rede de influências obtida.

médio da rede é 1.65, o que indica que a maior parte dos usuários postou ou foi citado poucas vezes durante o período. A distribuição de graus de entrada e de saída de G podem ser vistas na Figura 12. Ambas seguem uma lei de potência, com a grande maioria dos vértices com um grau baixo. Apesar disso, o grau de saída atinge valores muito maiores que o grau de entrada. Isso é reflexo de uma característica comum em redes sociais: um indivíduo é influenciado por alguns poucos, porém uma única celebridade, por exemplo, pode influenciar muitos outros usuários.

A assortatividade (*assortativity coefficient*) em relação ao grau é muito baixa, porém negativa: há maior tendência de conversa entre vértices de graus diferentes, o que é esperado em uma rede social como o *Twitter*, onde é comum citar personalidades famosas.

A Figura 13 mostra a evolução temporal do grau médio, *assortativity coefficient* e *clustering* à medida que mais arestas são adicionadas à rede. Podemos perceber que há uma tendência à estabilidade dessas medidas, o que nos faz concluir que continuar amostrando a rede durante mais tempo não deve alterar significativamente suas características principais.

Figura 12 – Distribuição de graus da rede de influências obtida. (a) Grau de Entrada (b) Grau de Saída

Figura 13 – Evolução temporal da rede. (a) *Clustering*. (b) Grau Médio. (c) *Mixing Assortativity*

4.2.2 Difusão de Hashtags

Paralelamente aos dados de menções e *retweets*, armazenamos, no mesmo período, informações sobre todas as *hashtags* compartilhadas fornecidas pela API do Twitter. A cada novo *tweet* que possui uma *hashtag*, armazenamos o usuário que a compartilhou, o nome da *hashtag* e a data completa do *post*.

Esses dados nos permitem montar, para cada *hashtag*, um conjunto ordenado ε de tuplas (u, t) , onde u é um usuário que compartilhou a *hashtag* e t o tempo. A Tabela 2 mostra as 10 *hashtags* mais populares no período, em termos de usuários únicos que a compartilharam. Dentre os temas que aparecem estão: política (*PiliPinasDebates2016*), música (*LEMONADE*, 방탄소년단, *GOT7*), TV (*GameofThrones*) e jogos (トレクル, モンストせいや). A Figura 14 mostra a distribuição de quantidade de usuários únicos por *hashtag*. Como esperado, há muitas *hashtags* compartilhadas por poucos, enquanto que há um grupo pequeno de *hashtags* bastante populares.

Estamos interessados em descobrir como a estrutura da rede influencia no processo de difusão dessas *hashtags*. Para responder essa pergunta, é interessante entender como evolui a popularidade de cada um desses tópicos ao longo do tempo. A Figura 15 mostra a distribuição acumulada ao longo do tempo da porcentagem do número total de usuários observados que compartilhou uma *hashtag*, para algumas das *hashtags* mais populares mostradas na Tabela 2.

Podemos distinguir, à partir da Figura 15, dois comportamentos principais. A evolução da *hashtag* トレクル (Figura 15 (c), uma *hashtag* sobre um jogo) é quase constante ao longo

<i>Hashtag</i>	#Usuários únicos
LEMONADE	17921
PiliPinasDebates2016	11186
トレクル	10643
ProjectHome	10266
GameofThrones	10134
방탄소년단	8882
gameinsight	8278
モンスターセいや	7875
GOT7	7623
RTした人全員フォローする	7576

Tabela 2 – Dez *hashtags* mais compartilhadas no período observado. Figura 14 – Distribuição de quantidade de usuários únicos por *hashtag*.

do tempo observado. Outros exemplos de *hashtags* que apresentam esse comportamento são: *#travel*, *#Trump*, *#job* e *#Venezuela*. Um tópico que apresenta uma constante evolução ao longo do tempo dificilmente representa um processo de difusão. Uma das características principais de processos de difusão é a existência de perda de interesse após uma evolução inicial, o que não é observado nesses casos. A maior parte das *hashtags* com taxas quase constantes de evolução temporal se referem a conceitos gerais, que são bastante utilizados, mas que não instigam um processo de discussão na rede social.

No extremo oposto à Figura 15 (c), temos o gráfico da Figura 15 (b), que é uma *hashtag* relacionada aos debates presidenciais nas Filipinas. Um número muito alto de *tweets* foram realizados durante o período do debate presidencial, e o interesse decaiu rapidamente logo em seguida. Podemos atribuir esse comportamento, geralmente, a alguma *novidade*: na *hashtag* *#LEMONADE*, temos um novo álbum da cantora *Beyoncé*; em *GameOfThrones*, um episódio novo da série televisiva *Game of Thrones*. Nosso interesse está, principalmente, no processo de difusão desse tipo de tópicos na rede social. Podemos notar que a grande maioria dos tópicos se refere a um evento que aconteceu *externamente* à rede, mas que rapidamente se propagou no *Twitter*. Estamos interessados em descobrir qual o papel da rede social nesse processo.

Figura 15 – Popularidade acumulada das 6 *hashtags* mais compartilhadas no período observado.

Para distinguir os dois comportamentos, propomos uma medida chamada *SWT* (*smallest time window*, ou menor janela de tempo). A $STW(p)$ é o tamanho da menor janela de tempo que contém pelo menos uma proporção, p (que é um parâmetro), das observações. Propomos um algoritmo de complexidade $O(N \log T)$, onde N é o número de observações e T o tamanho do intervalo de tempo das observações. O algoritmo é composto de duas etapas: uma busca binária no valor da resposta e uma checagem linear. A observação que nos permite aplicar a busca binária é a seguinte: se existe uma janela de tamanho t_m que contém pelo menos uma proporção p das observações, com certeza existe uma janela de tamanho $t > t_m$. Basta então descobrir se há uma janela que contenha a proporção p para um dado t fixo: isso pode ser feito utilizando um algoritmo de janelas deslizantes. Juntando as duas ideias, temos o Algoritmo 3.

A Figura 16 mostra alguns intervalos obtidos pelo Algoritmo 3 aplicado às *hashtags* com

Algoritmo 3: Algoritmo para calcular a menor janela de tempo que contém uma proporção p das observações

Input: E , sequência de pontos representando compartilhamentos de uma *hashtag*, ordenados

Input: Proporção p desejada

Output: $SWT(p)$

```

1  $S \leftarrow$  array de tamanho  $|E|$ 
2  $N \leftarrow |E|$ 
3  $S[0] \leftarrow E[0]$  for  $i \leftarrow 1 \dots N$  do
4    $S[i] \leftarrow S[i-1] + E[i]$ 
5  $N_e \leftarrow p \times |E|$ 
6  $lo \leftarrow 1$ 
7  $hi \leftarrow |E|$ 
8 while  $lo \neq hi$  do
9    $mi \leftarrow \frac{lo+hi}{2}$ 
10   $j \leftarrow 0$ 
11   $g \leftarrow False$ 
12  for  $i \leftarrow 0 \dots |E|$  do
13    while  $j < |E|$  and  $E[j] - E[i] \leq mi$  do
14       $j \leftarrow j + 1$ 
15      if  $j - i \geq N_e$  then
16         $g \leftarrow True$ 
17  if  $g = True$  then
18     $hi \leftarrow mi$ 
19  else
20     $lo \leftarrow mi + 1$ 
21 return  $lo$ 

```

a proporção de 80%. Observamos que, quanto mais constante é a taxa de compartilhamento da *hashtag*, mais próximo $SWT(p)$ será de p , o que nos permite diferenciar os comportamentos discutidos anteriormente. Utilizamos essa medida nas seções seguintes para caracterizar cascatas de informação.

4.3 Caracterização de Cascatas

Utilizando os métodos apresentados na Seção 4.1, construímos as cascatas de informação referentes a cada uma das *hashtags* presente no conjunto de dados. Para o restante da análise, utilizamos apenas as *hashtags* com mais de 1000 compartilhamentos. A Tabela 3 mostra o número de *hashtags* diferentes e compartilhamentos totais e para o grupo de *hashtags* com mais de 1000 ocorrências. Podemos notar que, apesar do grupo de tópicos compartilhado mais de 1000 vezes ser pequeno em relação ao conjunto total (0.05%), esses tópicos são responsáveis por (23.40%) dos compartilhamentos observados no período.

Figura 16 – SWT(80%) aplicada às séries das *hashtags* *GameOfThrones* e *トレクル*. A região destacada em cinza corresponde à menor janela encontrada.

Medida	Total	> 1000 compartilhamentos
Quantidade de <i>hashtags</i>	719095	429 (0.05%)
Quantidade de compartilhamentos	5170618	1205202 (23.40%)

Tabela 3 – Quantidade de *hashtags* observadas e compartilhamentos, totais e para *hashtags* com pelo menos 1000 compartilhamentos.

A Figura 17 exemplifica duas cascatas, formadas pelas *hashtags* #PiliPinasDebates2016 e #EarthDay. É possível observar uma forte presença de *hubs* periféricos, que reforçam a propagação de informação.

Figura 17 – Cascata formada pelas *hashtags* *PiliPinasDebates2016* e *EarthDay*. Apenas o maior componente conectado (considerando arestas bidirecionais) está visível.

Um grande número de compartilhamentos não necessariamente implica em uma cascata

<i>Hashtag</i>	#Usuários únicos	#Arestas na cascata
LEMONADE	17921	932
PiliPinasDebates2016	11186	6055
トレクル	10643	5
ProjectHome	10266	5436
GameofThrones	10134	1911
방탄소년단	8882	3348
gameinsight	8278	2
モンスターせいや	7875	19
GOT7	7623	5331
RTした人全員フォローする	7576	1406

Tabela 4 – Dez *hashtags* mais compartilhadas no período observado e tamanho de cascata relacionado.

com um grande número de arestas. A Tabela 4 mostra o número de arestas presentes nas cascatas referentes às *hashtags* da Tabela 2. Como discutido na Seção 4.2.2, a cascata referente à *hashtag* *トレクル*, por exemplo, que apresentou taxa quase constante de compartilhamento, apresentou um número muito baixo de arestas, o que é um indício forte de que os fatores principais para a propagação dessa informação são externos à rede.

Na Seção 4.1, discutimos sobre a probabilidade condicional de um vértice v fazer parte do processo de propagação de informação dado que ele é um vértice intermediário - está entre dois outros vértices que foram observados na amostra. Chamados essa probabilidade de P_i . Intuitivamente, quanto maior P_i , mais troca de informação interna na rede está acontecendo. Se considerarmos os seguintes eventos:

- E_p : v participa da cascata
- E_q : v é um vértice intermediário (existe um caminho $a \rightarrow v \rightarrow b$ tais que a e b foram amostrados na cascata)

Temos então que P_i pode ser calculada utilizando a relação da Equação 4.3:

$$P_i = P(E_p|E_q) = \frac{P(E_q \cap E_p)}{P(E_p)} \quad (4.3)$$

$P(E_p)$ nada mais é que a densidade da *hashtag*, isto é, a porcentagem do número de usuários que compartilhou a *hashtag* na rede. Para estimar P_i , precisamos então estimar $P(E_q \cap E_p)$. Para isso, calculamos a proporção dos vértices que foram observados na cascata que são também são vértices intermediários.

Para descobrir se um vértice é intermediário, varremos a lista de adjacência da cascata e construímos os vetores v_{in} e v_{out} . Um vértice v está em v_{in} se algum vértice x na cascata C possui a aresta $x \rightarrow v$. Da mesma forma, v está em v_{out} se existe algum y tal que exista a aresta $v \rightarrow y$. Consideramos v intermediário se ele está em v_{in} e também em v_{out} .

A Figura 18 mostra a distribuição das medidas P_i e $SWT(80\%)$ para o conjunto das 429 *hashtags* utilizadas. Na Figura 18 (b), podemos observar que as situações mais raras são as de grandes picos de compartilhamento ($SWT < 0.4$) e de taxa constante ($SWT > 0.7$).

Figura 18 – Distribuição de valores para P_i e $SWT(80\%)$ no conjunto de dados de 429 *hashtags*. As barras no eixo x representam a distribuição unidimensional dos dados.

Argumentamos que um maior valor de P_i deve implicar em um processo de difusão que está sendo bastante influenciado pela estrutura da rede. Da mesma forma, um valor menor de SWT representa uma propagação bastante rápida da informação. Esperamos que processos de propagação de informação em cascatas ocorram em intervalos de tempo pequenos, então espera-se que exista uma correlação entre as medidas. A Figura 19 (a) mostra a relação entre as duas medidas. O coeficiente de correlação de Pearson obtido foi de -0.19 - negativo, como o esperado. Na Figura 19 (b) temos a distribuição de SWT e a densidade (proporção de usuários únicos que compartilhou a *hashtag*). Vemos que não existe nenhuma relação entre a velocidade de propagação de um tópico e número total de usuários atingidos. O coeficiente de correlação, nesse caso, foi de 0.07 .

Figura 19 – Relação entre P_i , $SWT(80\%)$ e densidade.

Na Seção 4.1, discutimos outra medida de caracterização para cascatas de informações: o número de caminhos disjuntos para a informação, que chamamos de HC . Lembramos que HC é uma medida de quantas versões distintas da informação podem existir na cascata observada. Um valor alto implica que é possível que muitas das histórias observadas em vértices folhas (aqueles que não passam a informação adiante) sejam totalmente distintas. Um valor baixo significa que provavelmente existem alguns vértices onde a informação muito provavelmente deve passar para se propagar pela rede, que limitam a quantidade de histórias únicas possíveis. A Figura 20 (a) mostra a distribuição dos valores de HC no conjunto de dados observado. Como exemplo, o valor de HC para a rede da Figura 17 (b), referente a *hashtag PiliPinasDebates2016* foi de 433. Como observamos 11186 usuários compartilhando essa *hashtag*, podemos concluir que existe uma boa possibilidade de existirem muitas opiniões diferentes sobre esse debate. É importante notar que HC é uma medida de *potencial* de heterogeneidade: ela nos diz quantas versões independentes de uma história podem existir no máximo, o que não significa que a informação realmente seguiu caminhos disjuntos.

Figura 20 – Distribuição de HC para todas as *hashtags*.

Na Figura 21 podemos analisar como a medida HC se relaciona com P_i e SWT . Mais uma vez, notamos que um valor maior de HC é um sinal de que há um processo de propagação ocorrendo na rede, então é esperado que haja uma correlação positiva com P_i e negativa com SWT . Vemos que há uma forte correlação entre P_i e HC (0.87), muito maior que a relação entre P_i e SWT (0.19), mas ambas seguem os resultados esperados.

Figura 21 – Relações entre P_i , $SWT(80\%)$ e HC .

Uma outra medida relevante em cascatas de informação é a distância média que a informação trafega em relação a algum vértice fonte. Esperamos que esse número seja baixo em redes sociais, onde o número médio de pessoas entre quaisquer duas pessoas é baixo. A Figura 22 (a) mostra a distribuição da distância média nas cascatas observadas. Podemos observar que a grande maioria das cascatas observadas têm distância média menor que 1.0. A principal razão disso são, muito provavelmente, os vértices isolados, que não estão conectados a nenhum outro vértice mas mesmo assim compartilham a informação. a

Figura 22 – Distribuições da distância média e relação com o número de caminhos disjuntos.

Na Figura 22 (b) podemos observar que HC e a distância média estão bastante relacionadas. Isso é esperado, pois, para que existam diferentes versões de uma informação, é necessário que ela se propague para além da fonte.

4.4 Caminhos Mínimos e Centralidade

Até aqui, propusemos quatro medidas relevantes para caracterização de cascatas de informação e as estudamos no contexto da rede social *Twitter*.

Uma pergunta natural que surge em processos de propagação de informação é a seguinte: *a informação se propaga por caminhos mínimos?* Uma das medidas de centralidade mais

utilizadas em trabalhos científicos, a *betweenness centrality*, tem como principal ponto de partida essa suposição. Podemos utilizar cascatas de informação para investigar essa afirmação, uma vez que elas representam o processo de propagação de informação em uma rede.

Vamos considerar novamente uma cascata C inferida sobre uma rede G . Se conseguirmos calcular o menor caminho entre todos os pares de vértices em C e fazê-lo simultaneamente em G , podemos investigar se a propagação está, de fato, se propagando através de caminhos mínimos.

Vamos utilizar os dados obtidos nas seções anteriores para realizar essa estimativa. Para isso, consideramos o seguinte procedimento. Para cada cascata gerada por uma *hashtag*, calculamos o caminho mínimo entre os vértices fonte, com grau de entrada 0 e todos os outros vértices da cascata. Eliminamos vértices isolados e vértices inatingíveis. Calculamos então a mesma distância na rede de influência original. Para calcular caminhos mínimos, utilizamos uma simples busca em largura.

Obtivemos os caminhos mínimos entre 224662 pares de vértices no total. Calculamos então a diferença entre o caminho mínimo na rede original e na rede original. A distribuição dos valores da diferença pode ser vista na Figura 23

Figura 23 – Distribuição da diferença entre caminho mínimo e caminho obtido.

Podemos observar que a grande maioria dos pares seguiu o caminho mínimo. A média das diferenças foi de 0.42, com 75% dos pares observados em cascatas comunicando-se através do caminho mínimo. Se considerarmos que esse número é proveniente de amostras, se o caminho observado não for o real, ele é, com certeza, maior que o caminho real. Podemos afirmar, então, que sim: a informação trafega, principalmente, por caminhos mínimos, apesar de nem sempre. Esse efeito, entretanto, parece ser uma propriedade *emergente*, já que não há organização nenhuma nesse sentido.

Uma vez observada a importância dos caminhos mínimos no processo de propagação de

informação, podemos utilizar a *betweenness centrality* como medida de centralidade. Entretanto, o melhor algoritmo conhecido para o cálculo dessa medida é o apresentado em (BRANDES, 2001), cuja complexidade é de $O(N^2 \log N)$, onde N é o número de vértices da rede. Para uma rede com milhões de vértices, esse cálculo ainda é muito custoso.

Apesar disso, o algoritmo proposto por Brandes é paralelizável. O algoritmo, a cada etapa, monta a árvore de caminhos mínimos proveniente de cada vértice v e calcula as colaborações parciais de v ao cálculo do *betweenness centrality*. Não vamos detalhar o procedimento, pois ele pode ser encontrado na referência.

Para paralelização, podemos dividir o conjunto de vértices em K conjuntos de tamanho $\frac{N}{K}$, onde K é o número de processos que utilizaremos para o cálculo total. Cada um dos processos receberá um grupo de vértices para o qual deverá calcular as colaborações para todos os outros vértices. Devido a natureza do cálculo, podemos no final apenas combinar todos os resultados parciais em um único resultado final. O *speed-up* obtido, nesse caso, é linear com o número de processos utilizados, K .

Figura 24 – *Betweenness centrality* como função do Grau.

Para efetuar o cálculo da *betweenness centrality*, utilizamos o algoritmo de Brandes com 20 nós de processamento. A Figura 24 mostra o resultado da *betweenness centrality* em função do grau dos vértices na rede. Em uma rede social como o *Twitter*, o número direto de ligações (no nosso caso, citações) é um dos indicadores mais óbvios de influência. Notamos que, de fato, vértices com maior grau aparecem na metade superior dos valores de *betweenness*, porém a relação não é absoluta. Podemos observar alguns vales na distribuição, com os maiores valores de *betweenness* não correspondendo exatamente aos maiores graus da rede. Esse resultado indica que há uma diferença considerável entre a importância de um vértice para a propagação de informação e sua importância local, o número de outros vértices influenciados por ele diretamente.

4.5 Conclusão

Nesse capítulo, propusemos quatro medidas para caracterização de cascatas, visando, ao mesmo tempo, responder algumas perguntas fundamentais sobre o processo de propagação de informação em redes. Apresentamos dois algoritmos: para o cálculo do número de caminhos disjuntos em um *DAG* e para o cálculo da menor janela de tempo que contém uma proporção das observações de uma série. Ambos os algoritmos foram apresentados com sua complexidade esperada, e funcionaram bem no contexto dos conjuntos dados com os quais trabalhamos.

Realizamos uma análise quantitativa e qualitativa das medidas propostas aplicadas em dados reais. Concluimos que todas são eficazes em dizer se a propagação de uma informação está ocorrendo como um processo dinâmico na rede ou se está apenas sendo observada na rede. Além disso, apesar de correlacionadas, as medidas não são equivalentes, o que nos permite utilizá-las em conjunto.

Por fim, mostramos que a informação geralmente se propaga por caminhos mínimos no *Twitter*. Esperamos que isso se reflita em outras redes de informação, mas um estudo específico do domínio deve ser realizado utilizando a mesma metodologia. Desenvolvemos uma variação paralela de um conhecido algoritmo para o cálculo da *betweenness centrality* em uma rede com milhões de vértices.

Todos os resultados foram obtidos através de amostras. Um passo importante a ser dado num trabalho futuro é estudar um domínio com informação completa, e entender como o processo de amostragem afeta a inferência de cascatas de informação. No Capítulo 3 apresentamos alguns trabalhos nesse sentido, mas ainda são poucos.

HOMOFILIA E REDES DE TÓPICOS

Nesse capítulo, buscamos responder as perguntas apresentadas na introdução:

- É possível agrupar tópicos usando apenas a topologia de rede (conexões entre usuários) no *Twitter*?
- Qual a influência da homofilia nas conexões da rede?

Na Seção 5.1 apresentamos um método para agrupamento de tópicos utilizando a topologia da rede do *Twitter*. Em seguida, na Seção 5.2, derivamos uma rede a partir de interesses comuns e mostramos a importância da homofilia na formação das conexões na rede social.

5.1 Detecção de Tópicos

Com o intuito de entender como os tópicos compartilhados no *Twitter* estão relacionados entre si, representados nesse trabalho como *hashtags*, utilizamos os dados obtidos pela *API* de *streaming* descritos na Seção 2.1 para construir uma rede não direcionada de co-ocorrências, conectando duas *hashtags* sempre que as mesmas aparecem no mesmo *tweet*.

Um mesmo par de *hashtags* pode aparecer juntos em um *tweet* mais de uma vez, o que é um indício mais forte de que as duas estão relacionadas de alguma forma — geralmente tratam do mesmo tópico. Para incluir essa informação na rede de tópicos de co-ocorrência, podemos codificar a quantidade de co-ocorrências para um par de *hashtags* $\{a, b\}$ como o peso da aresta entre a e b na rede. Esse valor, porém, irá atribuir um peso muito maior para as *hashtags* que ocorrem mais vezes. Se chamarmos de W_{ab} o peso da aresta $\{a, b\}$, de T_a o conjunto de *tweets* que contém a *hashtag* a e T_b o conjunto de *tweets* que contém a *hashtag* b , resolvemos esse

problema fazendo

$$W_{ab} = \frac{|T_a \cap T_b|}{|T_a|} \quad (5.1)$$

Dessa forma, se duas *hashtags* sempre aparecem juntas, o peso da aresta entre as duas é 1, e se elas nunca aparecem juntas, o peso da aresta correspondente é 0. A rede obtida possui 719095 vértices (cada um correspondendo a uma *hashtag* diferente) e 4100274 arestas. Uma pequena amostra de aproximadamente 5% dessa rede pode ser vista na Figura 25.

Figura 25 – Pequena amostra da rede de co-ocorrências de *hashtags*. Cores correspondem à comunidades.

Em seguida, utilizamos o método de Louvain (BLONDEL *et al.*, 2008) para descobrir comunidades nessa rede, de forma a tentar agrupar as *hashtags* por tópicos. A modularidade obtida foi de 0.77 em duas iterações do método, o que é um valor bastante alto. Foram obtidas 299660 comunidades. Algumas das comunidades obtidas podem ser vistas na Tabela 5.

O alto valor de modularidade é um indício de que o método é eficaz para o agrupamento de *hashtags* em tópicos, e a inspeção manual de algumas comunidades obtidas está de acordo com esse resultado. Entretanto, uma análise mais rigorosa dessas partições ainda pode ser realizada como trabalho futuro, considerando, inclusive, o conteúdo escrito dos *tweets* e o significado semântico (que nem sempre é inteligível) das *hashtags*.

Tópico	Exemplos de <i>hashtags</i>
Política Brasileira	dilmashamesbrazil, TchauQuerida, ImpeachmentJa SomosTodosMoro, RindoMuito
Dia da Terra	EarthDay2016, PeopleCan, loveourplanet, savetheRhinos EarthDayFail, NaturalScenes, nofossil
The Voice	TheVoice, NeverForgerAnthonyTouma, TheVoiceTop11 GlobalGiftGala, VoterSlimane
Eleições Americanas	Bernie2016, BernieOrBastille, WeThePeople Referendum, NoHillary, VoteToStayInEurope

Tabela 5 – Exemplos de *hashtags* agrupadas na mesma comunidade.

5.2 Homofilia e Topologia de Rede no *Twitter*

Utilizando os tópicos obtidos na Seção 5.1, é possível determinar a amplitude da influência da homofilia na organização das conexões da rede. Tratamos aqui a homofilia como a presença de interesses comuns entre pessoas.

Partimos da rede de influências obtida na Seção 4.2.1. Nessa rede, a aresta $a \rightarrow b$ existe se o usuário b cita o usuário a ou o usuário b *retweeta* o usuário a . Chamaremos essa rede de G .

Derivamos então uma outra rede, G' , a partir de G . O conjunto de vértices de G' é o mesmo de G , porém as arestas são incluídas quando há interesses comuns entre vértices. Desejamos determinar um procedimento para descobrir as arestas de G' e medir o quão parecida é a configuração final com a configuração original de G .

De modo a selecionar quais serão as arestas presentes em G' , mantemos os graus de saída dos vértices, ou seja, o número de outros vértices por ele influenciado, iguais em G e G' . Dessa maneira, a tarefa de calcular G' se torna a tarefa de encontrar, para cada usuário u que possui $|V(u)|$ vizinhos em G , quem são os outros $|V(u)|$ usuários cujos interesses são os mais próximos possíveis dos interesses de u .

Utilizamos como medida de similaridade para um conjunto de interesses o índice de Jaccard entre dois conjuntos. Seja I_a o conjunto de tópicos compartilhados pelo usuário a , onde um tópico está representado pela classe de modularidade de uma *hashtag* compartilhada por a , e I_b o conjunto para um usuário b , temos que a similaridade de a e b , J_{ab} , é dada por

$$J_{ab} = \frac{|I_a \cap I_b|}{|I_a \cup I_b|} \quad (5.2)$$

A medida J_{ab} penaliza usuários que compartilham muitos assuntos. Para calcular o conjunto de arestas correspondente ao vértice a , todos os outros vértices b da rede são ranqueados com respeito a J_{ab} , e os $|V(a)|$ maiores valores são escolhidos como arestas. Esse procedimento, entretanto, tem complexidade $O(N^2 \log(|V(a)|))$ se implementado com uma *heap*, uma vez que analisamos todos os outros vértices para cada vértice da rede.

Nossa implementação realiza esse exato procedimento, porém, os assuntos compartilha-

dos por um grande número de usuários são ignorados no cálculo, o que permite que o algoritmo termine muito mais cedo. Assumimos aqui que, quanto mais popular um tópico, menos relevante é o interesse comum no mesmo para determinar a existência de uma aresta.

Para determinar o quão similar a rede G' é a G , propomos uma métrica baseada na semelhança entre as vizinhanças dos vértices de G e G' . Como as duas são redes com os mesmos vértices, podemos medir diretamente a diferença, D_g entre os conjuntos de arestas de um mesmo vértice em G e em G' . Para isso, calculamos a média do índice de Jaccard dos conjuntos de arestas vértice a vértice

$$D_g = \frac{1}{N} \sum_{i=1}^N \frac{|V_G(x) \cap V_{G'}(x)|}{|V_G(x) \cup V_{G'}(x)|} \quad (5.3)$$

Onde $V_G(x)$ é a vizinhança do vértice x em G e $V_{G'}(x)$ é a vizinhança do vértice x em G' . D_g varia de 0 a 1, onde um valor de 1 indica que G e G' são iguais. Aplicando o método na rede do *Twitter*, obtivemos um valor de $D_g = 0.025$.

Para decidir se esse valor é relevante, ou seja, se a homofilia tem relevância na formação das conexões, utilizamos o seguinte procedimento. Se assumirmos que as conexões não estão relacionadas aos tópicos compartilhados na rede, o processo de obtenção de G' se torna um processo randômico, onde cada vértice faz um *rewiring* de suas arestas para outros vértices sorteados aleatoriamente. Vamos chamar essa rede obtida nesse processo de G_r . Repetimos esse processo 1000 vezes, obtendo um D_g diferente para cada instância. A Figura 26 mostra a distribuição dos D_g obtidos. Podemos notar que a probabilidade de D_g ser 0.025 através de um processo de *rewiring* aleatório é praticamente nula - o menor valor obtido encontra-se na casa de 10^{-7} . Ou seja, obtivemos um índice de semelhança 10^5 vezes maior do que o esperado de um processo aleatório. Esse resultado nos permite concluir que a homofilia possui papel fundamental na organização das conexões da rede influenciadas do *Twitter*.

Figura 26 – Distribuição de D_g para várias amostras de G_r .

5.3 Conclusão

Nesse capítulo, mostramos que é possível utilizar somente a topologia da rede de co-ocorrências de *hashtags* para agrupá-las em tópicos com sucesso. Mostramos que essa rede possui uma alta modularidade, o que indica que existe uma estrutura de comunidades intrínseca na rede. Apesar disso, ainda é relevante e importante um trabalho mais detalhado sobre esse tema, e a exploração de medidas para definir o sucesso do agrupamento, além da própria modularidade.

Em seguida, mostramos que, ao reconstruir a rede de influência apenas utilizando a informação de tópicos compartilhados, conseguimos uma rede muito mais similar a original do que a que seria obtida através de um processo aleatório, o que nos permite concluir que a homofilia é, de fato, um fator de extrema relevância na construção das conexões na rede social.

CONCLUSÃO E TRABALHOS FUTUROS

Nesse trabalho, utilizamos conceitos de redes complexas e dados disponíveis da rede social *Twitter* para estudar o processo de difusão de informação e de criação de conexões nessa rede e em redes complexas em geral.

No Capítulo 4, desenvolvemos métodos e medidas para caracterização de cascatas, e conseguimos separar os processos de difusão de informação que são apenas observados na rede, porém ocorrem por fatores externos daqueles que se beneficiam mais diretamente da topologia da rede para difusão.

Os métodos e medidas desenvolvidos são relevantes para a caracterização de cascatas de informação. Apesar disso, um dos maiores problemas encontrados nessa etapa do trabalho é o caráter amostral dos dados. Propomos duas direções de trabalhos futuros que visam cobrir as falhas observadas nesse sentido: estudos teóricos sobre a influência da amostragem na observação de processos de difusão de informação e estudos com os mesmos métodos e medidas, porém com universos de dados completos ou modelos de difusão simulados. Há muito trabalho a ser feito nas duas direções sugeridas.

Também no Capítulo 4, mostramos que a informação de fato propaga-se principalmente por caminhos mínimos, com o auxílio dos dados de cascatas de informação obtidos do *Twitter*. Como trabalho futuro, é desejável um trabalho maior com mais fontes de dados, de forma a testar essa afirmação em outros domínios.

Por fim, no Capítulo 5 mostramos que a homofilia tem papel fundamental na construção das conexões de influência no *Twitter*, e mostramos que é possível agrupar tópicos utilizando apenas a estrutura da rede de co-ocorrência de *hashtags*. No entanto, nossa única medida de sucesso do método é a inspeção visual e a modularidade da rede, que é uma análise qualitativa. Em trabalhos futuros, podemos replicar os métodos utilizados em outros domínios, onde exista um agrupamento de tópicos que faça sentido semântico *a priori*, como na *Wikipedia*, para a qual já existem ontologias pré-compiladas. Dessa forma, seria possível determinar o sucesso do

método ao comparar as comunidades obtidas com as ontologias que já existem.

REFERÊNCIAS

- AIELLO, L. M.; BARRAT, A.; SCHIFANELLA, R.; CATTUTO, C.; MARKINES, B.; MENCZER, F. Friendship prediction and homophily in social media. **ACM Transactions on the Web (TWEB)**, ACM, v. 6, n. 2, p. 9, 2012. Citado na página 44.
- ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. **Reviews of modern physics**, APS, v. 74, n. 1, p. 47, 2002. Citado 2 vezes nas páginas 24 e 34.
- AMARAL, L.; OTTINO, J. Complex networks: Augmenting the framework for the study of complex systems. **The European Physical Journal B-Condensed Matter and Complex Systems**, Springer, v. 38, n. 2, p. 147–162, 2004. Citado na página 23.
- ANDERSON, R. M.; MAY, R. M. **Infectious Diseases of Humans Dynamics and Control**. [S.l.]: Oxford University Press, 1992. Citado na página 37.
- ARAL, S.; MUCHNIK, L.; SUNDARARAJAN, A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 106, n. 51, p. 21544–21549, 2009. Citado na página 44.
- BACKSTROM, L.; BOLDI, P.; ROSA, M.; UGANDER, J.; VIGNA, S. Four degrees of separation. In: ACM. **Proceedings of the 4th Annual ACM Web Science Conference**. [S.l.], 2012. p. 33–42. Citado na página 46.
- BAKSHY, E.; HOFMAN, J. M.; MASON, W. A.; WATTS, D. J. Everyone’s an influencer: Quantifying influence on twitter. Citeseer, 2011. Citado na página 24.
- BARABÁSI, A.-L. **Linked: How Everything is Connected to Everything Else and what it Means for Business, Science, and Everyday Life**. [S.l.]: Plume, 2003. Citado na página 24.
- BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **science**, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999. Citado 2 vezes nas páginas 24 e 35.
- BARRAT, A.; BARTHLEMY, M.; VESPIGNANI, A. **Dynamical processes on complex networks**. [S.l.]: Cambridge University Press New York, NY, USA, 2008. Citado 4 vezes nas páginas 36, 37, 38 e 39.
- BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **the Journal of machine Learning research**, JMLR. org, v. 3, p. 993–1022, 2003. Citado na página 42.
- BLONDEL, V. D.; GUILLAUME, J.-L.; LAMBIOTTE, R.; LEFEBVRE, E. Fast unfolding of communities in large networks. **Journal of Statistical Mecxhanics: Theory and Experiment**, IOP Publishing, v. 2008, n. 10, p. P10008, 2008. Citado na página 64.
- BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D.-U. Complex networks: Structure and dynamics. **Physics reports**, Elsevier, v. 424, n. 4, p. 175–308, 2006. Citado na página 29.

- BONDY, J. A.; MURTY, U. S. R. **Graph theory with applications**. [S.l.]: Macmillan London, 1976. v. 290. Citado na página 33.
- BRANDES, U. A faster algorithm for betweenness centrality*. **Journal of mathematical sociology**, Taylor & Francis, v. 25, n. 2, p. 163–177, 2001. Citado na página 61.
- CATALDI, M.; CARO, L. D.; SCHIFANELLA, C. Emerging topic detection on twitter based on temporal and social terms evaluation. In: ACM. **Proceedings of the Tenth International Workshop on Multimedia Data Mining**. [S.l.], 2010. p. 4. Citado na página 24.
- CHA, M.; HADDADI, H.; BENEVENUTO, F.; GUMMADI, P. K. Measuring user influence in twitter: The million follower fallacy. **ICWSM**, v. 10, p. 10–17, 2010. Citado na página 41.
- CHA, M.; MISLOVE, A.; GUMMADI, K. P. A measurement-driven analysis of information propagation in the flickr social network. In: ACM. **Proceedings of the 18th international conference on World wide web**. [S.l.], 2009. p. 721–730. Citado na página 43.
- COSTA, L.; Oliveira Jr, O.; TRAVIESO, G.; RODRIGUES, F.; BOAS, P.; ANTIQUEIRA, L.; VIANA, M.; Da Rocha, L. Analyzing and modeling real-world phenomena with complex networks: A survey of applications. **Advances in Physics**, v. 3, n. 60, p. 329–412, 2011. Citado na página 24.
- DALEY, D.; KENDALL, D. Epidemics and rumours. Nature Publishing Group, 1964. Citado na página 38.
- EDMONDS, J.; KARP, R. M. Theoretical improvements in algorithmic efficiency for network flow problems. **Journal of the ACM (JACM)**, ACM, v. 19, n. 2, p. 248–264, 1972. Citado na página 47.
- ERDÖS, P.; RÉNYI, A. On random graphs, i. **Publicationes Mathematicae (Debrecen)**, v. 6, p. 290–297, 1959. Citado na página 34.
- FIORE, A. T.; DONATH, J. S. Homophily in online dating: when do you like someone like yourself? In: ACM. **CHI’05 Extended Abstracts on Human Factors in Computing Systems**. [S.l.], 2005. p. 1371–1374. Citado na página 44.
- FORTUNATO, S. Community detection in graphs. **Physics Reports**, Elsevier, v. 486, n. 3, p. 75–174, 2010. Citado 2 vezes nas páginas 32 e 33.
- FREEMAN, L. C. Centrality in social networks conceptual clarification. **Social networks**, Elsevier, v. 1, n. 3, p. 215–239, 1978. Citado na página 34.
- GIRVAN, M.; NEWMAN, M. E. Community structure in social and biological networks. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 99, n. 12, p. 7821–7826, 2002. Citado 2 vezes nas páginas 31 e 32.
- GOEL, S.; WATTS, D. J.; GOLDSTEIN, D. G. The structure of online diffusion networks. In: ACM. **Proceedings of the 13th ACM conference on electronic commerce**. [S.l.], 2012. p. 623–638. Citado na página 43.
- GRANOVETTER, M. Threshold models of collective behavior. **American journal of sociology**, JSTOR, p. 1420–1443, 1978. Citado na página 42.

- HUBERMAN, B. A.; ROMERO, D. M.; WU, F. Social networks that matter: Twitter under the microscope. **Available at SSRN 1313405**, 2008. Citado na página 41.
- KANG, J. H.; LERMAN, K. Using lists to measure homophily on twitter. In: CITESEER. **AAAI Workshops**. [S.l.], 2012. Citado na página 42.
- KEMPE, D.; KLEINBERG, J.; TARDOS, É. Maximizing the spread of influence through a social network. In: ACM. **Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.], 2003. p. 137–146. Citado na página 43.
- KWAK, H.; LEE, C.; PARK, H.; MOON, S. What is twitter, a social network or a news media? In: ACM. **Proceedings of the 19th international conference on World wide web**. [S.l.], 2010. p. 591–600. Citado na página 41.
- LEBENSZTAYN, E.; MACHADO, F.; RODRÍGUEZ, P. Limit theorems for a general stochastic rumour model. **SIAM Journal on Applied Mathematics**, v. 71, n. 4, p. 1476–1486, 2011. Disponível em: <<http://epubs.siam.org/doi/abs/10.1137/100819588>>. Citado na página 39.
- LYNCH, C. Big data: How do your data grow? **Nature**, Nature Publishing Group, v. 455, n. 7209, p. 28–29, 2008. Citado na página 23.
- MANYIKA, J.; CHUI, M.; BROWN, B.; BUGHIN, J.; DOBBS, R.; ROXBURGH, C.; BYERS, A. H. Big data: The next frontier for innovation, competition, and productivity. 2011. Citado na página 23.
- MCAFEE, A.; BRYNJOLFSSON, E. Big data: the management revolution. **Harvard business review**, n. 90, p. 60–6, 2012. Citado na página 23.
- MCPHERSON, M.; SMITH-LOVIN, L.; COOK, J. M. Birds of a feather: Homophily in social networks. **Annual review of sociology**, JSTOR, p. 415–444, 2001. Citado na página 44.
- MILGRAM, S. The small world problem. **Psychology today**, New York, v. 2, n. 1, p. 60–67, 1967. Citado 2 vezes nas páginas 35 e 46.
- MITCHELL, M. Complex systems: Network thinking. **Artificial Intelligence**, Elsevier, v. 170, n. 18, p. 1194–1212, 2006. Citado na página 23.
- _____. **Complexity: A guided tour**. [S.l.]: Oxford University Press, 2009. Citado na página 24.
- MORENO, Y.; NEKOVEE, M.; PACHECO, A. Dynamics of rumor spreading in complex networks. **Physical Review E**, APS, v. 69, n. 6, p. 066130, Jun 2004. Disponível em: <<http://link.aps.org/doi/10.1103/PhysRevE.69.066130>>. Citado na página 39.
- MORSTATTER, F.; PFEFFER, J.; LIU, H.; CARLEY, K. M. Is the sample good enough? comparing data from twitter’s streaming api with twitter’s firehose. **arXiv preprint arXiv:1306.5204**, 2013. Citado na página 28.
- NEKOVEE, M.; MORENO, Y.; BIANCONI, G.; MARSILI, M. Theory of rumour spreading in complex social networks. **Physica A: Statistical Mechanics and its Applications**, v. 374, n. 1, p. 457–470, 2007. ISSN 0378-4371. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0378437106008090>>. Citado na página 39.
- NEWMAN, M. E. Mixing patterns in networks. **Physical Review E**, APS, v. 67, n. 2, p. 026126, 2003. Citado 2 vezes nas páginas 30 e 31.

_____. The structure and function of complex networks. **SIAM review**, SIAM, v. 45, n. 2, p. 167–256, 2003. Citado na página 35.

NEWMAN, M. E.; GIRVAN, M. Finding and evaluating community structure in networks. **Physical review E**, APS, v. 69, n. 2, p. 026113, 2004. Citado na página 32.

NEWMAN, M. E. J. **Networks: An Introduction**. [S.l.]: Oxford University Press, 2010. Citado 3 vezes nas páginas 29, 33 e 38.

PALLA, G.; DERÉNYI, I.; FARKAS, I.; VICSEK, T. Uncovering the overlapping community structure of complex networks in nature and society. **Nature**, Nature Publishing Group, v. 435, n. 7043, p. 814–818, 2005. Citado na página 32.

RODRIGUEZ, M. G.; LESKOVEC, J.; KRAUSE, A. Inferring networks of diffusion and influence. In: ACM. **Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.], 2010. p. 1019–1028. Citado na página 43.

ROGERS, E. M. **Diffusion of innovations**. [S.l.]: Simon and Schuster, 2010. Citado na página 44.

SADIKOV, E.; MEDINA, M.; LESKOVEC, J.; GARCIA-MOLINA, H. Correcting for missing data in information cascades. In: ACM. **Proceedings of the fourth ACM international conference on Web search and data mining**. [S.l.], 2011. p. 55–64. Citado 2 vezes nas páginas 43 e 45.

SHALIZI, C. R.; THOMAS, A. C. Homophily and contagion are generically confounded in observational social network studies. **Sociological methods & research**, Sage Publications, v. 40, n. 2, p. 211–239, 2011. Citado na página 44.

SIMON, H. A. A behavioral model of rational choice. **The quarterly journal of economics**, JSTOR, p. 99–118, 1955. Citado na página 23.

SUH, B.; HONG, L.; PIROLI, P.; CHI, E. H. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: IEEE. **Social computing (socialcom), 2010 ieee second international conference on**. [S.l.], 2010. p. 177–184. Citado na página 42.

WATTS, D. A twenty-first century science. **Nature**, v. 445, n. 7127, p. 489, 2007. Citado na página 23.

WATTS, D. J. A simple model of global cascades on random networks. **PNAS**, v. 99, n. 9, 2002. Citado 2 vezes nas páginas 43 e 45.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of ‘small-world’ networks. **nature**, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998. Citado 2 vezes nas páginas 31 e 35.

WELCH, M. J.; SCHONFELD, U.; HE, D.; CHO, J. Topical semantics of twitter links. In: ACM. **Proceedings of the fourth ACM international conference on Web search and data mining**. [S.l.], 2011. p. 327–336. Citado na página 42.

WENG, L. **Information diffusion on online social networks**. Tese (Doutorado), 2014. Citado na página 42.

WENG, L.; LENTO, T. M. Topic-based clusters in egocentric networks on facebook. In: **ICWSM**. [S.l.: s.n.], 2014. Citado na página 44.

YAN, G.; ZHOU, T.; HU, B.; FU, Z.-Q.; WANG, B.-H. Efficient routing on complex networks. **Physical Review E**, APS, v. 73, n. 4, p. 046108, 2006. Citado na página [33](#).

ZACHARY, W. W. An information flow model for conflict and fission in small groups. **Journal of anthropological research**, JSTOR, p. 452–473, 1977. Citado na página [23](#).

ZAMAN, T. R.; HERBRICH, R.; GAEL, J. V.; STERN, D. Predicting information spreading in twitter. In: CITESEER. **Workshop on computational social science and the wisdom of crowds, nips**. [S.l.], 2010. v. 104, n. 45, p. 17599–601. Citado na página [42](#).

ZHAO, W. X.; JIANG, J.; WENG, J.; HE, J.; LIM, E.-P.; YAN, H.; LI, X. Comparing twitter and traditional media using topic models. In: **Advances in Information Retrieval**. [S.l.]: Springer, 2011. p. 338–349. Citado na página [42](#).