***Dinâmicas de propagação de
informações e rumores em redes
sociais***

Didier Augusto Vega Oliveros
Tese de Doutorado

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Didier Augusto Vega Oliveros

Dinâmicas de propagação de informações e rumores em redes sociais

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Francisco Aparecido Rodrigues

USP – São Carlos
Julho de 2017

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

V422d Vega Oliveros, Didier Augusto
Dinâmicas de propagação de informações e rumores em
redes sociais / Didier Augusto Vega Oliveros;
orientador Francisco Aparecido Rodrigues. -- São
Carlos, 2017.
170 p.

Tese (Doutorado - Programa de Pós-Graduação em
Ciências de Computação e Matemática Computacional) --
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2017.

1. Redes sociais. 2. Propagação de informação. 3.
Propagação de rumores. 4. Problema de maximização da
influência. 5. Redes complexas. I. Rodrigues,
Francisco Aparecido , orient. II. Título.

Didier Augusto Vega Oliveros

Information and rumor propagation in social networks

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Francisco Aparecido Rodrigues

USP – São Carlos
July 2017

AGRADECIMENTOS

A minha linda e querida esposa Lilian, pelo amor e carinho, pelo grande apoio, compreensão e ajuda em todos os momentos. Com você, a minha vida é cheia de alegrias!. *También, quiero agradecer a mi familia por el ánimo, la buena energía y sus oraciones.*

A meus orientadores, da graduação Profa. Zoila Inés Ramos de Flores, mestrado Profa. Maria da Graça Campos Pimentel e no doutorado, Prof. Luciano da Fontoura Costa e especialmente ao Prof. Francisco Aparecido Rodrigues, pela paciência e apoio durante estes anos e toda a ajuda oferecida. A todos, muito obrigado pelos valiosos ensinamentos.

Also, I would like to acknowledge and thank Dr. Albert-László Barabási and Dr. Federico Vázquez, to each one for the opportunity to visit his research group and the valuable contributions to my researcher career.

Quero agradecer a todos os professores que contribuíram na minha formação para poder atingir este objetivo, desde o colégio, a *Universidad Nacional de Colombia, Universidad de los Andes* e do Instituto de Ciências Matemáticas e de Computação (ICMC-USP). *Especialmente, quiero agradecer a los profesores Miguel Peña, Dra. Angela Cristina Carrillo Ramos y Dr. Dave Machado, quienes me incentivaron y ayudaron en momentos claves.*

Agradeço aos colegas do ICMC e franciscanos, a meus colaboradores como Dr. Zhao Liang, Dr. Antonio Mendez-Bermudez, Dr. Alneu de Andrade Lopes, entre outros não menos importantes. Aos amigos, tanto aqui no Brasil como fora, pelos bons momentos compartilhados, o bom humor e paciência. A todos os funcionários do ICMC, à USP e ao Brasil, pelo apoio financeiro concedido mediante a CNPq e por ter-me adotado como filho.

Finalmente, agradeço a Deus pelo presente da vida.

*“O fato de uma opinião ser amplamente compartilhada
não é nenhuma evidência de que não seja completamente absurda;
de fato, tendo-se em vista a maioria da humanidade,
é mais provável que uma opinião difundida seja tola do que sensata.”*
(B. RUSSELL ([RUSSELL, 1970](#)))

RESUMO

VEGA-OLIVEROS, D. A. **Dinâmicas de propagação de informações e rumores em redes sociais**. 2017. 170 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

As redes sociais se tornaram um novo e importante meio de intercâmbio de informações, ideias e comunicação que aproximam parentes e amigos sem importar as distâncias. Dada a natureza aberta da Internet, as informações podem fluir muito fácil e rápido na população. A rede pode ser representada como um grafo, onde os indivíduos ou organizações são o conjunto de vértices e os relacionamentos ou conexões entre os vértices são o conjunto de arestas. Além disso, as redes sociais representam intrinsecamente a estrutura de um sistema mais complexo que é a sociedade. Estas estruturas estão relacionadas com as características dos indivíduos. Por exemplo, os indivíduos mais populares são aqueles com maior número de conexões. Em particular, é aceito que a estrutura da rede pode afetar a forma como a informação se propaga nas redes sociais. No entanto, ainda não está claro como a estrutura influencia na propagação, como medir seu impacto e quais as possíveis estratégias para controlar o processo de difusão. Nesta tese buscamos contribuir nas análises da interação entre as dinâmicas de propagação de informações e rumores e a estrutura da rede. Propomos um modelo de propagação mais realista considerando a heterogeneidade dos indivíduos na transmissão de ideias ou informações. Nós confirmamos a presença de propagadores mais influentes na dinâmica de rumor e observamos que é possível melhorar ou reduzir expressivamente a difusão de uma informação ao selecionar uma fração muito pequena de propagadores influentes. No caso em que se objetiva selecionar um conjunto de propagadores iniciais que maximizem a difusão de informação, a melhor opção é selecionar os indivíduos mais centrais ou importantes nas comunidades. Porém, se o padrão de conexão dos vértices está negativamente correlacionado, a melhor alternativa é escolher entre os indivíduos mais centrais de toda a rede. Por outro lado, através de abordagens topológicas e de técnicas de aprendizagem máquina, identificamos aos propagadores menos influentes e mostramos que eles atuam como um firewall no processo de difusão. Nós propomos um método adaptativo de reconexão entre os vértices menos influentes para um indivíduo central da rede, sem afetar a distribuição de grau da rede. Aplicando o nosso método em uma pequena fração de propagadores menos influentes, observamos um aumento importante na capacidade de propagação desses vértices e da rede toda. Nossos resultados vêm de uma ampla gama de simulações em conjuntos de dados artificiais e do mundo real e a comparação com modelos clássicos de propagação da literatura. A propagação da informação em redes é de grande relevância para as áreas de publicidade e marketing, educação, campanhas políticas ou de saúde, entre outras. Os resultados desta tese podem ser aplicados e estendidos em diferentes campos de pesquisa como redes biológicas e modelos de comportamento social animal, modelos de propagação de epidemias e

na saúde pública, entre outros.

Palavras-chave: Redes sociais, Propagação de informação, Propagação de rumores, Problema de maximização da influência, Redes complexas.

ABSTRACT

VEGA-OLIVEROS, D. A. **Information and rumor propagation in social networks**. 2017. 170 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

On-line Social networks become a new and important medium of exchange of information, ideas and communication that approximate relatives and friends no matter the distances. Given the open nature of the Internet, the information can flow very easy and fast in the population. The network can be represented as a graph, where individuals or organizations are the set of vertices and the relationship or connection among the vertices are the set of edge. Moreover, the social networks are also intrinsically representing the structure of a more complex system that is the society. These structures are related with characteristics of the subjects, like the most popular individuals have many connections, the correlation in the connectivity of vertices that is a trace of homophily phenomenon, among many others. In particular, it is well accepted that the structure of the network can affect the way the information propagates on the social networks. However, how the structure impacts in the propagation, how to measure that impact and what are the strategies for controlling the propagation of some information, it is still unclear. In this thesis, we seek to contribute in the analysis of the interplay between the dynamics of information and rumor spreading and the structure of the networks. We propose a more realistic propagation model considering the heterogeneity of the individuals in the transmission of ideas or information. We confirm the presence of influential spreaders in the rumor propagation process and found that selecting a very small fraction of influential spreaders, it is possible to expressively improve or reduce the diffusion of some information on the network. In the case we want to select a set of initial spreaders that maximize the information diffusion on the network, the simple and best alternative is to select the most central or important individuals from the network's communities. But, if the pattern of connection of the networks is negatively correlated, the best alternative is to choose from the most central individuals in the whole network. On the other hand, we identify, by topological approach and machine learning techniques, the least influential spreaders and show that they act as a firewall in the propagation process. We propose an adaptive method that rewires one edge for a given vertex to a central individual, without affecting the overall distribution of connection. Applying our proposed method in a little fraction of least influential spreaders, we observed an important increasing in the capacity of propagation of these vertices and in the overall network. Our results are from a wide range of simulations in artificial and real-world data sets and the comparison with the classical rumor propagation model. The propagation of information is of greatest relevance for publicity and marketing area, education, political or health campaigns, among others. The results of this these might be applicable and extended in different research fields like biological networks and animal social

behavior models.

Keywords: Social Networks, Information spreading, Rumor propagation, Influence maximization problem, Complex networks.

LISTA DE ILUSTRAÇÕES

Figura 1	– Distribuição de grau de 30 iterações para uma rede ER com $N = 500$, $P = 0.045$, $\langle k \rangle = 23.76 \pm 0.33$, e $\langle \ell \rangle = 2.223 \pm 0.10$	41
Figura 2	– Modelo de Watts-Strogatz para redes mundo pequeno. Adaptada de Watts e Strogatz (1998a)	42
Figura 3	– Característica do modelo WS para rede com $N = 600$: (a) e (b), distribuição de grau segundo P ; (c) média do coeficiente de <i>clustering</i> $\langle C \rangle$ VS a média dos menores caminhos geodésicos $\langle g \rangle$ com $m = 6$ para 30 iterações	43
Figura 4	– Modelo de rede sem escala Barabási e Albert: (a) exemplo de uma rede BA com $N = 500$, $m_0 = m = 3$ e $\gamma = 2$; (b) distribuição de grau para 30 simulações com os mesmos parâmetros e resultado de $\langle k \rangle = 5.976 \pm 0.0$, $\langle \ell \rangle = 2.749 \pm 0.048$ e $\langle C \rangle = 0.215 \pm 0.026$	45
Figura 5	– Análise de centralidade para a rede Dolphins (LUSSEAU <i>et al.</i> , 2003): (a) quanto mais vermelho escuro maior o grau e quanto maior o vértice maior seu grau de intermediação; (b) quanto mais azul escuro menor <i>eccentricity</i> e quanto maior o vértice maior o valor do <i>PageRank</i>	49
Figura 6	– Exemplo de decomposição em núcleos de uma rede. (a) representa o conjunto de vértices mais centrais com $Kc = 3$; (b) vértices no núcleo $Kc = 2$; (c) vértices no núcleo $Kc = 1$	53
Figura 7	– Dendograma das comunidades encontradas pelo algoritmo de <i>fastgreedy</i> para a rede clube de karatê. As formas dos vértices representam os dois grupos nos quais o clube se separou.	59
Figura 8	– Arcabouço da transmissão de uma doença. O sistema inclui: (1) os possíveis hospedeiros, (2) o patógeno ou agente infeccioso, (3) os fatores que estimulam a proliferação, (4) as interações de contato entre os indivíduos, e (5) a dinâmica interna da doença nos indivíduos infectados.	61
Figura 9	– Resposta temporal dos infectados para a dinâmica SI em epidemias, nos modelos de redes ER, WS e BA com $N = 1000$, $\langle k \rangle = 8$, $\beta = 0.1$ e $\mu = 0.2$	66
Figura 10	– Resposta temporal para a dinâmica SIS em epidemias, nos modelos de redes ER, WS e BA com $N = 1000$, $\langle k \rangle = 8$, $\beta = 0.1$ e $\mu = 0.2$. A curva vermelha representa a fração de indivíduos infectados e a azul os suscetíveis.	69
Figura 11	– Resposta temporal para a dinâmica SIR em epidemias, nos modelos de redes ER, WS e BA com $N = 1000$, $\langle k \rangle = 8$, $\beta = 0.1$, $\mu = 0.2$, e as respectivas curvas da fração de infectados (vermelho), suscetíveis (azul) e recuperados (verde)	72

Figura 12 – Caso homogêneo. Propagação de rumor para uma rede Barabási-Albert (BA) com tamanho de $N = 10^4$ vértices: (a) densidade final de contidos (vermelho = 1, azul = 0) para diferentes taxas de propagação (linhas brancas tracejadas). Valores iguais de λ produzem a mesma densidade esperada de contidos; (b) comparação dos resultados teóricos (linha contínua) com simulações de Monte Carlo (MC) (símbolos). As figuras internas mostram a probabilidade de que o vértice x ($x = 1, \dots, N$) seja um contido (R_x , pontos vermelhos) ou um ignorante (S_x , pontos verdes) para simulações de MC e resultados teóricos. Cada simulação é uma média de mais de 500 realizações. 98

Figura 13 – Caso heterogêneo. Densidade final de contidos em um rede Barabási-Albert (BA) de tamanho $N = 10^4$. Os símbolos correspondem as simulações de Monte Carlo (MC) enquanto que as linhas contínuas são soluções da análise teórica. A probabilidade de propagação β_x é proporcional às medidas de centralidade: (a) grau, (b) intermediação e (c) PageRank. As figuras internas mostram a probabilidade de que o vértice x ($x = 1, \dots, N$) seja um contido (R_x , pontos vermelhos) ou um ignorante (S_x , pontos verdes) para as simulações de MC e previsões teóricas. Cada simulação é uma média de 500 realizações. 99

Figura 14 – Densidade final de contidos de acordo com o nível de heterogeneidade definido pela função de interpolação (Eq. 4.10). As medidas de centralidade consideradas são centralidade de intermediação (BE), centralidade de proximidade (CL), grau (DG), centralidade de autovetor (EV), k -core (KC) e PageRank (PR). A propagação homogênea é dada para $p = 1$, enquanto que o caso completamente heterogêneo ocorre para $p = 0$. As simulações são realizadas considerando modelos de redes (a) Erdős-Rényi (ER), (b) Watts-Strogatz (WS), (c) Barabási-Albert (BA) e (d) espaciais sem escala de Barthélemy (SSF). A taxa de propagação $\lambda = \sqrt{2}$ foi adotada em todas as simulações. 102

Figura 15 – Densidade final de contidos em função da proporção de trocas aleatórias da probabilidade de propagação β_x nas redes geradas pelos modelos: (a) Erdős-Rényi (ER), (b) Watts-Strogatz (WS), (c) Barabási-Albert (BA) e (d) espaciais sem escala de Barthélemy (SSF). As medidas de centralidade empregadas são a centralidade de intermediação (BE), a centralidade de proximidade (CL), o grau (DG), centralidade de autovetor (EV), k -core (KC) e PageRank (PR). 103

Figura 16 – Fração final de contidos de acordo com o nível de heterogeneidade definido pela função de interpolação (Eq. 4.10) nas redes sociais: (a) *netscience*, (b) *email*, (c) *hamsterster* e (d) *Google+*. As medidas de centralidade consideradas são: centralidade de intermediação (BE), centralidade de proximidade (CL), grau (DG), centralidade de autovetor (EV), k -core (KC) e PageRank (PR). A propagação homogênea é dada para $p = 1$, enquanto que o caso completamente heterogêneo ocorre para $p = 0$ 104

Figura 17 – Diagramas de fase do comportamento da propagação ao longo do tempo na rede <i>Google++</i> . Considerou-se a abordagem de propagação clássica com β_x homogêneo. Foi melhorada ou controlada a capacidade de propagação selecionando um grupo de $\tau = 0.5\%$ vértices alvo e mudando suas probabilidades de propagação para um novo β_x^τ . A seleção dos τ vértices foi dos que tinha maior centralidade de PageRank. Também foi feita a seleção uniformemente aleatória de τ vértices para melhorar (Iprv) ou controlar (Ctrl) a difusão. (a) diagrama de fase dos propagadores ao longo do tempo normalizado pelo pico máximo da abordagem clássica; (b) diagrama de fase da capacidade de propagação final ao longo do tempo normalizado pela capacidade de propagação final da abordagem clássica.	107
Figura 18 – Modelo Maki e Thompson (1973) para propagação de rumores em uma rede artificial BA de $N = 1000$ vértices, $\langle k \rangle = 8$ e $\eta = 1\% V $ de sementes iniciais. A barra de cor simboliza a fração final de contidos: (a) processo reativo (RP); (b) processo truncado (TP); e (c) processo de contato (CP)	115
Figura 19 – Modelo MT-TP em uma rede sem escala artificial com $N = 1000$ e $\langle k \rangle = 8$. A fração final de indivíduos informados é definida nas barras de cor. A seleção de $\eta = 4\% V $ de sementes foi conforme: (a) o método proposto baseado em comunidades considerando os vértices com melhor centralidade de PageRank; (b) selecção aleatória uniforme de η vértices; e (c) os η nós com a maior centralidade PR da rede. As linhas brancas contínuas à esquerda representam combinações de parâmetros (β, μ) que atingiram 35% de indivíduos contidos. As linhas brancas tracejadas mostram as combinações que alcançaram 60% de indivíduos informados.	115
Figura 20 – Capacidade de propagação do modelo MT-TP nas quatro redes reais consideradas, dada a seleção de sementes por comunidades (pontos vermelhos), melhor ranqueados ou centrais (pontos pretos) e aleatoriamente (pontos azuis).	117
Figura 21 – Representação gráfica circular de uma rede BA com a correlação de graus modificada pela abordagem de Xulvi-Brunet e Sokolov (2005): (a) BA disassortiva com $\rho = -0.43$ e (b) BA assortativa com $\rho = 0,34$	123
Figura 22 – Métodos aqui considerados para tratar do problema de maximização da influência (IMP).	126
Figura 23 – Impacto da correlação grau-grau no problema de maximização de influência com $ S $ no intervalo de 2 a $90\% V $ de vértices. Para cada rede sem escala artificial calculamos o conjunto de sementes de acordo com: (BST) os vértices melhor classificados da rede; (COM) os vértices mais centrais das comunidades; (RANDOM) selecionando aleatoriamente as sementes; e (GREEDY) o método guloso de otimização. As medidas de centralidade adotadas são a centralidade de intermediação (<i>BE</i>), o grau (<i>DG</i>) e o PageRank (<i>PR</i>).	127

Figura 24 – Impacto da correlação grau-grau no pico de vértices informantes. O número de sementes está entre 2 a 10% do tamanho da rede (as setas indicam o aumento de η), selecionando as sementes pelo método guloso (<i>GREEDY</i>), comunidades (<i>COM</i>) e melhores ranqueados segundo a centralidade (<i>BST</i>).	129
Figura 25 – Velocidade do pico de vértices informantes no IMP de acordo com a correlação de grau. O número de sementes η está na faixa de 2 a 10% $ V $ do tamanho da rede, selecionado pelos métodos: (a) melhores classificações (<i>BST</i>), (b) comunidades (<i>COM</i>) e (c) de otimização guloso <i>GREEDY</i>	130
Figura 26 – Impacto da correlação de grau nas redes sociais. O número de sementes varia de 2 a 10 % do tamanho da rede, selecionado de acordo com as comunidades (<i>COM</i>), ranking dos nós mais centrais (best-rank, <i>BST</i>), aleatório (<i>RANDOM</i>) e pelo método de otimização gulosa (<i>GREEDY</i>). A centralidade de intermediação (<i>BE</i>), grau (<i>DG</i>) e PageRank (<i>PR</i>) são as medidas de centralidade utilizadas.	131
Figura 27 – Velocidade de propagação para atingir o pico de propagadores. O número de sementes varia de 2 a 10% $ V $, selecionando-os de acordo com os métodos: de otimização gulosa (<i>GREEDY</i>), comunidades (<i>COM</i>) e melhores classificados na rede (<i>BST</i>).	132
Figura 28 – A diferença crítica (CD) para comparar a classificação média de métodos diferentes no 95 percentil é 1.66, de acordo com o teste de Nemenyi. Diferenças significativas acima deste valor são significativas e estão desconectadas.	135
Figura 29 – Comparativo de z-score da capacidade de propagação em função da proporção de propagadores menos influentes.	142
Figura 30 – Análise de propagação para as redes sociais reais <i>hamsterster</i> , <i>advogado</i> e <i>Google+</i> : (a-c) o impacto de vértices com baixa centralidade no processo de propagação; (d-f) similaridade entre os vértices com baixa centralidade e os menos influentes; (g-i) características topológicas dos propagadores menos influentes e os vértices com os menores valores de <i>EV</i> . As simulações para <i>hamsterster</i> e <i>advogado</i> foram considerados os valores de $\beta = 0.3$ e $\mu = 0.1$, e para <i>Google+</i> $\beta = 0.5$ e $\mu = 0.1$	143
Figura 31 – Análise de propagação para as redes sociais ao realizar a proposta adaptativa trocando uma aresta dos 0.5% dos propagadores menos influentes, e conectando-os com os vértices com maior grau da rede: o impacto dos valores de baixa centralidade na média de propagação.	147

LISTA DE TABELAS

Tabela 1	– Propriedades estruturais das redes complexas analisadas.	99
Tabela 2	– Capacidade de propagação final de acordo com as centralidades	106
Tabela 3	– Propriedades estruturais e resultados da detecção de comunidades nas redes adotadas. O melhor valor de modularidade Q e o número de comunidades dessa divisão, segundo o algoritmo fastgreedy, estão na última coluna	113
Tabela 4	– Média das capacidades de propagação para os valores de $\eta \in [1, 50]$, para cada rede social e método de seleção: (segunda coluna maior) selecionando os vértices mais centrais das comunidades; (Terceira coluna maior), selecionando os vértices mais centrais da rede; e (última coluna maior) selecionando-os aleatoriamente. Em negrito destacamos os melhores resultados entre as métricas. As medidas adotadas foram a centralidade de intermediação (BE), grau (DG) e PageRank (PR).	117
Tabela 5	– Propriedades estruturais das redes consideradas. ρ é o coeficiente de assortatividade, $ V $ o tamanho da rede, $\langle k \rangle$ o grau médio, $\langle \ell \rangle$ é o comprimento médio dos caminhos mínimos e $\langle C_i \rangle$ e a média do coeficiente de agrupamento. Os parâmetros relacionados à comunidade são a modularidade Q e o número de comunidades N_c	123
Tabela 6	– A variação normalizada da informação (NVI) entre a distribuição de comunidades e as sementes GREEDY e BST-DG.	133
Tabela 7	– Resultados da influência ($\sigma(S)$) dos métodos analisados.	134
Tabela 8	– Propriedades estruturais das redes complexas	141
Tabela 9	– Resultados do método de agrupamento k -means para a rede <i>Google+</i>	144
Tabela 10	– Melhoria na capacidade global de propagação para as redes sociais: Primeiros lugares estão em negrito e segundos lugares estão sublinhados.	147

LISTA DE ABREVIATURAS E SIGLAS

J	Coeficiente de Jaccard
AN	Rede assortativa ou com correlação positiva na distribuição de grau
BA	Modelo de rede sem escala Barabási e Albert (1999)
BE	Centralidade de intermediação, do inglês <i>Betweenness centrality</i> (FREEMAN, 1977)
CL	Centralidade de proximidade, do inglês <i>Closeness centrality</i> (SABIDUSSI, 1996)
CP	Processo de contato, do inglês <i>Contact Process</i>
CT	Coeficiente de aglomeração, do inglês <i>Clustering coefficient</i> (WATTS; STROGATZ, 1998a)
DG	Centralidade de grau ou conectividade, do inglês <i>Degree centrality</i>
DK	Modelo Daley e Kendall (1964) de rumores
DN	Rede dissassortativa ou com correlação negativa na distribuição de grau
DP	Dinâmica de teoria de jogos do dilema do prisioneiro
ER	Modelo de rede Erdős e Rényi (1959)
EV	Centralidade de autovetor, do inglês <i>Eigenvector centrality</i> (BONACICH, 1972)
HO	Centralidade de buracos estruturais, do inglês <i>Structural holes</i> (BURT, 1992)
Homofilia	Homofilia, do inglês <i>homophily</i> , é uma teoria em sociologia que diz que as pessoas tendem a se relacionar com outros que sejam semelhantes a elas em características como status socioeconômico, valores, crenças ou atitudes (WENG; MENCZER; AHN, 2013 ; LOPES; PINTO; FRANCISCO, 2016).
IC	Modelo de cascata independente para propagação de informação (KEMPE; KLEINBERG; TARDOS, 2003), do inglês <i>Independent Cascade model</i>
IMP	Problema de maximização da influência ou da propagação de informação, do inglês <i>Influence Maximization Problem</i>
KC	Centralidade de <i>k-core</i> , do inglês <i>K-core centrality</i> (BATAGELJ; ZAVERSNIK, 2003)
LT	Modelo de limiar linear para propagação de informação (KEMPE; KLEINBERG; TARDOS, 2003), do inglês <i>Linear Threshold model</i>
MT	Modelo Maki e Thompson (1973) para rumores
NC	Rede não correlacionada ou sem correlação na distribuição de grau
PR	Centralidade de PageRank, do inglês <i>PageRank centrality</i> (BRIN; PAGE, 1998)
RP	Processo reativo, do inglês <i>Reactive Process</i>

SSF	Modelo de rede espacial sem escala (BARTHÉLEMY, 2003), do inglês <i>spatial scale-free model</i>
TP	Processo truncado, do inglês <i>Truncated Process</i>
WS	Modelo de rede Watts e Strogatz (1998a)

LISTA DE SÍMBOLOS

G — Rede

V — Conjunto de vértices

E — Conjunto de arestas

W — Conjunto de pesos nas arestas

N — Número de vértices

M — Número de arestas

A — Matriz de adjacência

A_{ij} — Elemento (i, j) da matriz de adjacência

$K(i)$ — Conjunto de vizinhos de i

G_r — Rede r -Regular

ℓ_{ij} — Comprimento do menor caminho entre os vértices i e j

ℓ_{max} — Diâmetro da rede

Cp — Principal componente da rede

P — Probabilidade de conexão entre dois vértices

k_i — Vértice i com grau k

$P(k)$ — Distribuição de probabilidade para k

B_i — Centralidade de *betweenness* do vértice i

C_t — Coeficiente de transitividade da rede

C_i — Centralidade de *Clustering* do vértice i

$C\ell_i$ — Centralidade de *Closeness* do vértice i

X_i — Centralidade de *Eigenvector* do vértice i

π_i — Centralidade de *PageRank* do vértice i

ρ — Correlação de Pearson para o grau da rede (Assortatividade)

ρ_s — Correlação de Spearman

β — Taxa de contágio ou difusão

μ — Taxa de recuperação ou contenção

S — Estado de suscetíveis ou desinformados do rumor

I — Estado de informantes do rumor, do inglês *informer* ou *spreader*

R — Estado de contidos do rumor, do inglês *reserved* ou *stifler*

SUMÁRIO

1	INTRODUÇÃO	27
1.1	Motivação e contextualização	28
1.2	Lacunas da área	30
1.3	Objetivos e trabalhos de pesquisa propostos	31
1.3.1	<i>Novos modelos de propagação de informação</i>	31
1.3.2	<i>Influência da estrutura da rede no processo de propagação</i>	32
1.3.3	<i>Técnicas para controle da propagação</i>	33
1.4	Contribuições	33
1.5	Estrutura do documento	35
2	FUNDAMENTAÇÃO TEÓRICA	37
2.1	Redes complexas	37
2.1.1	<i>Conceitos básicos</i>	38
2.1.2	<i>Modelos de redes artificiais</i>	39
2.1.2.1	<i>Modelos de redes aleatórias</i>	40
2.1.2.2	<i>Modelos de redes de pequeno mundo</i>	42
2.1.2.3	<i>Modelos de redes sem escala</i>	44
2.1.2.4	<i>Modelo de rede sem escala espacial</i>	45
2.1.3	<i>Medidas de centralidade</i>	46
2.1.3.1	<i>Centralidade de grau (DG)</i>	47
2.1.3.2	<i>Grau de intermediação (BE)</i>	47
2.1.3.3	<i>Medidas de vizinhança</i>	48
2.1.3.4	<i>Centralidade de proximidade (CL)</i>	50
2.1.3.5	<i>Centralidade de autovetor (EV)</i>	51
2.1.3.6	<i>Centralidade PageRank (PR)</i>	52
2.1.3.7	<i>Centralidade K-core (KC)</i>	52
2.1.3.8	<i>Centralidade de buracos estruturais (HO)</i>	53
2.1.4	<i>Medidas globais</i>	54
2.1.4.1	<i>Medidas relacionadas a conectividade</i>	54
2.1.4.2	<i>Medidas de distância</i>	56
2.1.4.3	<i>Modularidade</i>	57
2.2	Modelos de Propagação	60
2.2.1	<i>Propagação de epidemias</i>	60

2.2.1.1	<i>Abordagem geral</i>	62
2.2.1.2	<i>Modelagem matemática</i>	63
2.2.1.3	<i>Influência da estrutura da rede</i>	71
2.2.2	<i>Propagação de rumores</i>	74
2.2.2.1	<i>Abordagem geral</i>	75
2.2.2.2	<i>Influência da estrutura</i>	77
2.2.2.3	<i>Modelos estendidos</i>	79
2.2.2.4	<i>Modelagem matemática</i>	81
3	ABORDAGEM GERAL	85
3.1	Modelo de difusão adotado	85
3.2	Configuração das simulações	87
3.3	Redes complexas consideradas	89
4	MODELO HETEROGÊNEO DE PROPAGAÇÃO DE RUMORES . .	93
4.1	Motivação	94
4.2	Trabalhos relacionados	94
4.3	Proposta do modelo heterogêneo	95
4.4	Redes consideradas	98
4.5	Metodologia	100
4.6	Resultados	100
4.6.1	<i>Redes artificiais</i>	<i>100</i>
4.6.2	<i>Redes sociais</i>	<i>101</i>
4.7	Cenários de estudo: melhora e controle da propagação	103
4.7.1	<i>Cenário de controle da propagação de desinformação</i>	<i>104</i>
4.7.2	<i>Cenário de melhora na propagação da informação</i>	<i>105</i>
4.7.3	<i>Resultados dos cenários de estudo</i>	<i>105</i>
4.8	Considerações finais	108
5	SELEÇÃO DOS VÉRTICES MAIS INFLUENTES NAS COMUNIDADES	109
5.1	Motivação	109
5.2	Trabalhos relacionados	110
5.3	Seleção de propagadores influentes segundo as comunidades	112
5.4	Redes consideradas	113
5.5	Metodologia	113
5.6	Resultados	114
5.6.1	<i>Impacto do comportamento dos propagadores</i>	<i>114</i>
5.6.2	<i>Análises da seleção de múltiplas sementes</i>	<i>116</i>

5.7	Considerações finais	118
6	IMPACTO DA ASSORTATIVIDADE NA MAXIMIZAÇÃO DE IN- FORMAÇÃO	119
6.1	Motivação	120
6.2	O problema de maximização de informação	121
6.3	Redes consideradas	122
6.4	Metodologia	124
6.5	Resultados	127
6.5.1	<i>Impacto da assortatividade em redes artificiais</i>	127
6.5.2	<i>Impacto da assortatividade em redes reais</i>	129
6.5.3	<i>Distribuição das sementes</i>	132
6.6	Análise estatística	135
6.7	Considerações finais	137
7	IDENTIFICAÇÃO DOS PROPAGADORES MENOS INFLUENTES DA REDE	139
7.1	Motivação	140
7.2	Definição dos vértices menos influentes	141
7.3	Redes consideradas	141
7.4	Os propagadores menos influentes	141
7.4.1	<i>Análise topológica</i>	142
7.4.2	<i>Análise não supervisionada</i>	144
7.5	Método adaptativo para melhorar a propagação de informação	145
7.6	Resultados do método adaptativo	146
7.7	Considerações finais	148
8	CONCLUSÕES	149
8.1	Resultados e contribuições	151
8.1.1	<i>Contribuições</i>	151
8.1.2	<i>Publicações</i>	151
8.2	Limitações	153
8.3	Trabalhos Futuros	155
8.3.1	<i>Relação entre a estrutura e a dinâmica</i>	155
8.3.2	<i>Redes ponderadas</i>	155
8.3.3	<i>Dinâmica na população</i>	155
	REFERÊNCIAS	157

INTRODUÇÃO

Diversos sistemas complexos, entre eles a sociedade, meios de transporte como as linhas aéreas ou redes rodoviárias, meios de comunicação como a Internet e redes celulares, apesar de serem completamente distintos, apresentam uma característica em comum: são compostos por partes discretas que interagem de forma não-linear (BOCCALETTI *et al.*, 2006; COSTA *et al.*, 2011). A representação da estrutura de tais sistemas por meio de redes tem revelado propriedades universais, tais como um alto grau de heterogeneidade na organização das conexões e estrutura modular (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008; NEWMAN, 2010). Além disso, é possível modelar processos dinâmicos em tais estruturas, tais como propagação de informação (VESPIGNANI, 2011; CASTELLANO, 2012), rumores (ZANETTE, 2001a; MORENO; NEKOVEE; PACHECO, 2004; NEKOVEE *et al.*, 2007; BORGE-HOLTHOEFER; MORENO, 2012), cultura (AXELROD, 1997; KUPERMAN, 2006; APOLLONI; GARGIULO, 2011), epidemias (ANDERSON; MAY, 1992; KOOPMAN, 2005; KITSACK *et al.*, 2010), entre outros. Modelos que descrevem a evolução de uma epidemia em uma sociedade podem ser adaptados para se analisar a propagação de vírus na Internet ou de inovação tecnológica entre cidades (CASTELLANO; FORTUNATO; LORETO, 2009). Nesses casos, a organização das redes tem uma importância fundamental sobre a evolução dos processos dinâmicos (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008; CASTELLANO; FORTUNATO; LORETO, 2009).

Como um meio para estudar a organização e dinâmica em redes sociais, biológicas, tecnológicas e de informação, foi introduzida no final da década de 90 a Teoria das Redes Complexas (WATTS; STROGATZ, 1998a). Exemplos que podem ser representados por redes complexas são a *World Wide Web* (BARABÁSI, 2001; NEWMAN, 2003), a Internet (NEWMAN, 2003), redes sociais (GIRVAN; NEWMAN, 2002; KEELING; EAMES, 2005; BAKSHY *et al.*, 2012), redes de citações entre artigos científicos (NEWMAN, 2006), redes de interação de proteínas (ALBERT; BARABÁSI, 2002), entre muitas outras (COSTA *et al.*, 2011). Devido à generalidade das ferramentas utilizadas na análise de tais redes, a teoria das redes complexas apresenta um caráter altamente multidisciplinar, sendo utilizada nas mais diversas áreas, incluindo

Física, Sociologia, Ciência da Computação, Biologia, Epidemiologia e Engenharia (COSTA *et al.*, 2011).

Um dos principais estudos em redes objetiva entender os mecanismos em que opiniões, informações, vírus e rumores espalham-se sobre sua estrutura. Modelos de propagação podem ser usados no estudo de muitos sistemas sociais e físicos, que suportam a difusão de um sinal (KUPERMAN, 2006; FUNK; SALATHÉ; JANSEN, 2010; VIEIRA; SENNA; PEREIRA, 2011; GONZÁLEZ-BAILÓN; Borge-Holthoefer, Javier Rivero; MORENO, 2011; BAKSHY *et al.*, 2012).

1.1 Motivação e contextualização

Os estudos de processos de propagação em redes complexas têm servido para ajudar a explicar vários fenômenos de disseminação (ZANETTE, 2001a; NEWMAN, 2002b; BOCCALETTI *et al.*, 2006; BARABÁSI, 2007; BARRAT; BARTHÉLEMY; VESPIGNANI, 2008; CASTELLANO; FORTUNATO; LORETO, 2009), como por exemplo:

- Propagação de doenças ou vírus entre organismos vivos.
- Espalhamento de *softwares* maliciosos (*e.g. worms* ou vírus) através da Internet.
- Difusão do interesse por um produto ou moda, como música, filmes, estilo de roupas, APPs, entre outros.
- Espalhamento de uma ideia, conceito, cultura, notícia, memes ou senso comum popular.
- Propagação de falhas em materiais ou sistemas, tais como efeito dominó em um sistema elétrico.

Um exemplo particular de propagação foi a rápida adoção do jogo Pokémon GO (WINGFIELD, July, 2016; TABACCHI *et al.*, 2017; TEDXZURICH, 2017) lançado pela Niantic em julho de 2016 e que é baseado no vídeo jogo Pokémon da década de 90. O jogo tornou-se um fenômeno mundial com uma alta popularidade em pouco tempo, sendo o aplicativo da App Store iOS com mais downloads em uma semana do que qualquer outro App no mesmo período e um dos tópicos mais comentados em redes sociais, revistas on-line e tradicionais, entre outras mídias (TABACCHI *et al.*, 2017; TEDXZURICH, 2017). O sucesso deve-se em parte a uma estratégia que promove uma mistura interessante de realidade virtual e atividades de socialização em espaços públicos e redes sociais onde os jogadores podem interagir. A probabilidade de se tornar um jogador aumenta se mais vizinhos ou amigos estiverem jogando, convertendo-se em um efeito viral em cascata.

As abordagens propostas de modelos de propagação, tanto em epidemiologia (GRENFELL, 1997; KOOPMAN, 2002) ou sociologia (KEMPE; KLEINBERG; TARDOS, 2003;

RICHARDSON; DOMINGOS, 2002), indicam que existe um limiar ou ponto crítico para que aconteça uma difusão significativa através da rede. As informações ou doenças que se propagam em menor taxa a esse limite crítico tendem a desaparecer. Porém, as que se difundem acima desse limite se espalham exponencialmente e influenciam uma fração significativa da rede. Segundo a dinâmica e o tipo de rede é possível calcular o limite de propagação crítico. Por exemplo, para o caso de redes sociais ou tecnológicas, esse limite é próximo de zero (VESPIGNANI, 2011; BARRAT; BARTHÉLEMY; VESPIGNANI, 2008). Isso significa que qualquer vírus, inclusive aqueles com capacidade de contágio baixa, se mantém latentes na rede (BARABÁSI, 2007). Isso explica porquê vírus de computadores antigos ainda continuarem remanescente na Internet depois de anos da suposta erradicação.

Assim, existem indícios de que o processo de propagação está fortemente influenciado pela estrutura da rede (KEELING; EAMES, 2005; CASTELLANO; FORTUNATO; LORETO, 2009; CHEN *et al.*, 2013; MILLER; KISS, 2014). Características topológicas como a distribuição de grau, proporção de triângulos, correlação dos vértices, presença de *hubs*, entre outros, podem influenciar na capacidade de difusão de informação ou rumores. Foi verificado que uma série dessas características estruturais estão correlacionadas com traços da personalidade dos indivíduos (VISWANATH *et al.*, 2009; NOË; WHITAKER; ALLEN, 2016; TABACCHI *et al.*, 2017). O comportamento das pessoas instintivamente contribui para a organização das redes sociais, e em particular, a assortividade, a tendência de indivíduos semelhantes estar conectados, é uma propriedade central de homofilia nas redes sociais (VISWANATH *et al.*, 2009; TABACCHI *et al.*, 2017). Por exemplo, segundo os cinco fatores da personalidade¹ (NOË; WHITAKER; ALLEN, 2016; TABACCHI *et al.*, 2017), ser extrovertido e amável tem um efeito positivo no tamanho da rede do indivíduo, enquanto que o neuroticismo tem um efeito contrário. A extroversão e a abertura a novas experiências estão positivamente correlacionadas à transitividade e pessoas mais autodisciplinadas e cooperantes tendem a estar mais fortemente conectadas nos seus círculos de amizade. Uma possível explicação poderia ser que a personalidade, sendo uma característica inerente do indivíduo, desempenha um papel na determinação de sua posição na rede, e que por sua vez, a posição na rede reforça traços específicos da personalidade.

Analisar o impacto dessas estruturas locais e as dinâmicas de propagação permitem entender melhor como um “agente” se propaga e propor estratégias eficientes para controlar ou melhorar sua propagação (KITSACK *et al.*, 2010). Por exemplo, vacinar todas as pessoas pode ser muito custoso e a vacinação aleatória de indivíduos também não resolveria o problema. Porém, ao serem vacinados primeiro os indivíduos com maior número de conexões da rede, a imunização pode ser efetiva usando apenas uma pequena fração da população (BARABÁSI, 2007). Assim, ações focadas nos vértices mais fortemente conectados da rede podem ser vantajosas para imunização contra doenças, mas não necessariamente seria a melhor estratégia quando se busca a propagação massiva de informações. A partir do estudo de estratégias de propagação para redes,

¹ <[https://pt.wikipedia.org/wiki/Big_Five_\(psicologia\)](https://pt.wikipedia.org/wiki/Big_Five_(psicologia))>

podem-se planejar campanhas publicitárias que atinjam o maior número de pessoas no menor tempo e custo possível. Também, podem ser empregadas para que uma informação chegue de modo mais fácil e direto a um público alvo, diminuir ou anular o efeito de outra informação ou convencer o maior número de pessoas sobre alguma ideia, como no caso de campanhas publicitárias ou políticas.

Portanto, pesquisas no tema de propagação de informação em redes complexas são atuais e relevantes, pois permitem entender, evitar, controlar ou estimular a difusão de informação.

1.2 Lacunas da área

Apesar dos avanços obtidos na análise, modelagem e simulação de redes complexas, várias questões ainda precisam ser investigadas com maior profundidade (BRITTON *et al.*, 2015; PELLIS *et al.*, 2015). No caso das dinâmicas sociais, tais como propagação de informação e rumores, muitos modelos foram desenvolvidos (ANDERSON; MAY, 1992; BARRAT; BARTHÉLEMY; VESPIGNANI, 2008; DRAIEF; MASSOULI, 2010; PASTOR-SATORRAS *et al.*, 2015). No entanto, tais modelos são bastante simplificados e não consideram situações observadas em sistemas complexos reais. Ademais, o que se tem estudado sobre os fenômenos de propagação de epidemias e rumores foi feito considerando-se grafos totalmente conectados, sendo que apenas a partir de 2001 (PASTOR-SATORRAS; VESPIGNANI, 2001) passou-se a considerar modelos de redes com topologias ditas complexas. Tais redes ainda são bastante simplificadas, já que todos os vértices são semelhantes e, na grande maioria dos casos, também são estáticas.

Por outro lado, ainda não foi explorado em profundidade a presença de vértices influentes nos processos de propagação. De acordo com a dinâmica, podem existir vértices melhores propagadores, como no caso de epidemias (KITSACK *et al.*, 2010), ou não existir prova local significativa de existência, como no caso de rumores (BORGE-HOLTHOEFER; MORENO, 2012). Algumas pesquisas destacam o papel importante que possui a correlação de grau no processo de propagação (NEWMAN, 2002a; ná; PASTOR-SATORRAS; VESPIGNANI, 2003; NEKOVEE *et al.*, 2007). Porém, as análises não avaliam o nível de presença da assortatividade ou de outras propriedades topológicas sobre a rede e nem são consideradas abordagens em meso escala considerando a estrutura que permitam a identificação de grupos de melhores propagadores (KEELING; EAMES, 2005; PASTOR-SATORRAS *et al.*, 2015).

A inclusão de parâmetros que se adaptam junto com a dinâmica social, cujos vértices variam as conexões ou o potencial de propagação/recuperação ao longo do tempo tem sido pouco explorada (PASTOR-SATORRAS *et al.*, 2015; BRITTON *et al.*, 2015). Além disso, considerar vértices com taxas heterogêneas que podem ser modificadas na medida em que acontece a dinâmica, adição de parâmetros que considerem fenômenos reais, desenvolvimento de estratégias como desmerecer um rumor ou o surgimento de novas versões, competição, cooperação ou mutação das informações, entre outros, são temas em aberto na literatura (PASTOR-SATORRAS

et al., 2015; KEMPE; KLEINBERG; TARDOS, 2015; BRITTON *et al.*, 2015; PELLIS *et al.*, 2015). A evolução temporal, o limiar de propagação, a velocidade e pico de infectados, entre muitas outras propriedades dinâmicas são questões importantes a serem analisadas nas novas abordagens, de modo a verificar como a inclusão de propriedades mais realistas influencia na propagação.

Desse modo, existe a necessidade de modelos mais realistas, estudando a dinâmica de propagação em conjunto com os diferentes tipos de redes, de modo a verificar como a estrutura influencia a dinâmica. Nesse sentido, mecanismos de controle mais precisos sobre o impacto da propagação podem ser propostos. Logo, o campo de estudo de processos de propagação em rumores e epidemias para redes complexas ainda encontra-se em sua fase inicial e há muitas questões de pesquisa a serem exploradas.

1.3 Objetivos e trabalhos de pesquisa propostos

Este trabalho de doutorado teve como objetivo geral complementar os estudos realizados sobre propagação de informação e de rumores como dinâmicas sociais em redes. Objetivamos desenvolver modelos mais realistas, considerando: vértices com taxas de transmissão heterogêneas que podem ser modificados na medida em que acontece a dinâmica; adição de novos parâmetros que considerem fenômenos reais; adaptação da conexão dos vértices visando um incremento na capacidade de propagação da rede; desenvolvimento de estratégias como desmerecer um rumor para controlá-lo ou potencializar sua difusão. Também, buscou-se realizar estudos analíticos de modo a obter as equações que governam a dinâmica do modelo e verificar como os seus parâmetros influenciam na disseminação do rumor. Estudos de estabilidade e impacto dos parâmetros de tais modelos também são considerados.

A hipótese levantada neste trabalho considera que é possível desenvolver modelos de propagação mais realistas ao explorar diferentes medidas e propriedades topológicas da rede. Consideramos que a estrutura da rede exerce uma influência importante na dinâmica social, mas também que a dinâmica oferece uma retroalimentação para a rede, definindo a estrutura.

O objetivo geral pode ser dividido nos seguintes subprojetos específicos, que discutimos a seguir.

1.3.1 Novos modelos de propagação de informação

Novos modelos, que tenham propriedades mais próximas a realidade, foram propostos neste trabalho. Para isso foram avaliados fatores como: (i) taxas heterogêneas de propagação ou disposição para propagar, influências externas a rede, e (ii) abordagens adaptativas dos vértices. Modelos cujos vértices possuem probabilidades de propagação ou de contenção próprias visam analisar como a heterogeneidade afeta a dinâmica. Nesse sentido, abordamos as taxas

heterogêneas sendo estas dependentes a propriedades estruturais da rede ou independentes da estrutura, de acordo com uma distribuição de probabilidade.

O fator de adaptação da rede implica em que os vértices, sabendo seu estado e características de seus vizinhos, podem aumentar ou diminuir a força da conexão, inclusive, desfazer e criar novas conexões. Por exemplo, as pessoas tendem a evitar o contato com indivíduos que estejam manifestando algum sintoma de doenças transmissíveis, como a gripe ou tuberculose. Para todas as propostas, será analisado o nível de dependência do modelo com a topologia da rede, a fim de entender a relação entre a estrutura e a dinâmica.

1.3.2 Influência da estrutura da rede no processo de propagação

Podemos quantificar o nível de influência que a estrutura da rede exerce no processo de propagação. As análises foram considerando o impacto em escala local, como identificação de propagadores mais influentes; meso escala, como grupos de vértices que maximizem a propagação na rede; e global, quantificando o impacto que propriedades estruturais específicas têm sobre a dinâmica.

Trabalhos recentes têm considerado o impacto da estrutura na propagação (LI *et al.*, 2014), como por exemplo, a existência de vértices mais influentes em redes sem escala (MORENO; NEKOVEE; PACHECO, 2004; KITSACK *et al.*, 2010; BERGE-HOLTHOEFER; MORENO, 2012). No caso do modelo clássico de rumor, não existem vértices que sejam significativamente melhores que outros na propagação (BERGE-HOLTHOEFER; MORENO, 2012). Porém, a presença de alguma propriedade topológica na rede pode permitir a presença dos melhores propagadores. Desse modo, considerando uma propriedade topológica θ , podemos obter redes com diferentes valores dessa medida. Com isso, buscamos analisar o impacto que as propriedades estruturais têm sobre a dinâmica, identificar a presença de vértices melhores propagadores, em escala local, e verificar a influência na fração final de informados, em escala global. Especificamente para a análise em escala global, consideramos a diferença potencial da fração final de informados segundo cada propriedade, com $\Delta\theta_{xy} = \|c.v(\theta_x) - c.v(\theta_y)\|$, sendo θ_x o valor x que apresentou a rede para a medida θ . A função $c.v = \frac{\sigma(\theta_x)}{\langle\theta_x\rangle}$ é o coeficiente de variação da propriedade segundo o valor x , avaliando a média $\langle\theta_x\rangle$ e desvio padrão $\sigma(\theta_x)$ da fração de vértices informados.

Por outro lado, as abordagens para quantificar a relação entre a estrutura e a dinâmica consideram uma medida topológica. Assim, fizemos uma análise multivariável mediante um vetor de características estruturais para cada vértice i , com $\vec{v}_i = [\mathcal{F}_{1i}, \mathcal{F}_{2i}, \mathcal{F}_{3i}, \dots, \mathcal{F}_{ji}]$, onde $\mathcal{F}_{ji}$ é uma medida topológica do vértice i . A partir desses vetores de características pode ser empregado um método de mineração de dados para encontrar grupos de vértices, e desse modo, analisar em meso escala, o desempenho dos grupos em relação a capacidade de propagação.

1.3.3 Técnicas para controle da propagação

Desenvolvidos os modelos, podemos elaborar técnicas para controle da difusão. Por exemplo, estratégias de vacinação ou campanhas de prevenção no caso de processos de epidemia e estratégias publicitárias para alvos específicos ou de maior abrangência no caso de processos de difusão de informação. A partir do entendimento da relação entre a estrutura e a dinâmica, as técnicas propostas serão mais precisas e dependentes com a dinâmica e o tipo de rede. Trabalhos recentes identificaram uma relação de dependência em redes sem escala para a propagação de epidemias (KITSACK *et al.*, 2010). No entanto, outros tipos de topologia ainda não foram considerados, bem como redes espaciais. Desse modo, consideraremos diferentes tipos de redes e compararemos os resultados com abordagens desenvolvidas anteriormente, de modo a complementar ou generalizar técnicas existentes, ou desenvolver novas abordagens. Os métodos podem ser baseados em informações locais, onde não se possui ou conhece toda a rede, ou globais, que são usados quando se conhece toda a rede. Tais análises serão desenvolvidas nos modelos tradicionais e nos modelos propostos, aplicando-os em redes artificiais e reais.

1.4 Contribuições

Uma das contribuições foi a proposta de um modelo discreto de propagação de rumores com probabilidades heterogêneas de transmissão ou recuperação (VEGA-OLIVEROS; COSTA; RODRIGUES, 2017b). Cada vértice da rede possui sua própria probabilidade de contagiar a seus vizinhos ou de parar de propagar a informação. A partir dos resultados desse modelo foi proposto um método para anular ou melhorar a propagação de uma informação na rede, o qual melhorou em 35% e conteve em 60% a difusão da informação (VEGA-OLIVEROS; COSTA; RODRIGUES, 2017b). Por outro lado, foi feita uma análise em meso escala para identificar o grupo de vértices que mais propagam na rede (VEGA-OLIVEROS; BERTON, 2015). Para este problema, propusemos um método de identificação de sementes baseado na estrutura de comunidades (VEGA-OLIVEROS; BERTON, 2015). Verificamos o impacto que a correlação de grau da rede exerce na propagação de informação (VEGA-OLIVEROS; COSTA; RODRIGUES, 2017a). Os resultados do método proposto foram satisfatórios em comparação com outros métodos relevantes da literatura (VEGA-OLIVEROS; COSTA; RODRIGUES, 2017a). Finalmente, definimos e identificamos os vértices menos influentes da rede (VEGA-OLIVEROS *et al.*, 2015). Esses vértices têm sido desconsiderados nos trabalhos de propagação da literatura. Porém encontramos que é possível, mediante técnicas adaptativas, melhorar a capacidade de propagação desses vértices e da rede toda.

As simulações foram feitas em modelos de redes complexas onde foi possível controlar algumas propriedades, de modo a verificar como o comportamento segundo diferentes estruturas influencia no processo de propagação. Ademais, foi comparada a evolução temporal das abordagens propostas em tais redes com os modelos clássicos de propagação. Também, foram

empregadas redes de dados de domínios reais, tais como de redes sociais, blogs, Internet, de citações, entre outras.

Na investigação realizada neste doutorado, obtivemos as seguintes contribuições originais:

- Propusemos um modelo Markoviano discreto de propagação de rumores, o qual considera vértices com probabilidades de propagação e recuperação próprias (VEGA-OLIVEROS; COSTA; RODRIGUES, 2017b). Para o modelo proposto foi realizada a correspondente análise comparativa entre os resultados teóricos com os resultados das simulações, encontrando uma boa correspondência a nível local e global. Foi verificado como os parâmetros do modelo afetam o processo de propagação. Identificamos que ao adicionar probabilidades heterogêneas de transmissão surgem vértices mais influentes na dinâmica de rumor. Assim, vértices com maior grau, centralidade de intermediação ou PageRank propagam mais que os que apresentam maiores valores em outras centralidades. Também, avaliamos o modelo em diferentes topologias, tais como redes sem escala, pequeno mundo, espaciais, com distribuição de Poisson e regulares, bem como em dados reais, tais como dados do Google+, de troca de e-mails, redes tecnológicas, como mapas da Internet, entre outras. Como resultado, observamos que as taxas heterogêneas impactam mais nas redes com distribuição de grau heterogêneo, isto é com distribuição que segue uma lei de potência do que em redes regulares ou homogêneas.
- A partir dos resultados do modelo heterogêneo, observamos que os vértices com maior centralidade PageRank são propagadores influentes capazes de afetar a dinâmica de propagação, podendo aumentar ou reduzir a força da difusão (VEGA-OLIVEROS; COSTA; RODRIGUES, 2017b). Com uma pequena fração desses propagadores influentes (0.5% dos vértices com maior PageRank) é possível controlar a difusão de um rumor ou incrementar seu potencial de propagação entre 60% e 35% respectivamente. Observamos que o método proposto é mais efetivo no começo do processo de difusão, antes que 10% da população seja afetada. Os resultados mostram que técnicas que consideram a topologia da rede afetam o processo de propagação.
- Propusemos um método que seleciona um conjunto de η propagadores iniciais baseado na estrutura de comunidades da rede para maximizar a propagação de informação (VEGA-OLIVEROS; BERTON, 2015). O método proposto tem custo computacional baixo ($\mathcal{O}(n \log(n))$) o qual não depende do tamanho de η sementes. O método mais conceituado na literatura é um algoritmo guloso que apresenta custo computacional NP -difícil e que depende do número de η sementes requeridas. Mediante exaustivas simulações, em 6 redes artificiais e 8 redes reais, mostramos que não existe diferença estatística nos resultados entre o método proposto e o método guloso (VEGA-OLIVEROS; COSTA; RODRIGUES, 2017a).

- Avaliamos o impacto que a correlação de grau-grau tem sobre a dinâmica de propagação de informação em redes complexas (VEGA-OLIVEROS; COSTA; RODRIGUES, 2017a). Observamos que em redes disassortativas, ao selecionar um grupo de η sementes que sejam os vértices mais centrais da rede, é possível atingir um pico de indivíduos informados muito maior em comparação com outros métodos. No caso que a rede seja assortativa ou não assortativa, selecionar os vértices mais centrais não é a melhor opção, pois existe uma sobreposição de influência entre essas sementes. A medida que cresce o número de sementes iniciais η é possível que a fração final de informados diminua em comparação a η menores, o qual é um resultado não esperado. Desse modo, os modelos de propagação são funções monotônicas crescentes em relação a η quando as sementes são escolhidas de forma aleatória uniforme. Assim, métodos de maximização que não consideram a estrutura da rede, podem obter resultados subótimos.
- Identificamos, mediante análises topológica e usando o algoritmo k -means de aprendizado não supervisionado, os vértices que menos propagam na rede, os quais são aqueles com a menor centralidade de autovetor (VEGA-OLIVEROS *et al.*, 2015). Denominamos esses vértices como os vértices menos influentes ou piores propagadores. Não necessariamente os vértices que menos propagam informação são aqueles com grau ou centralidade de intermediação baixa na rede. Vértices com valores de centralidade acima da média da rede também podem ser propagadores menos influentes.
- Propusemos uma técnica adaptativa para aumentar a média de propagação dos vértices (VEGA-OLIVEROS *et al.*, 2015). A abordagem adaptativa é feita nos vértices menos influentes da rede. Ao fazer a reconexão de uma pequena fração desses vértices (0.5%) para os de maior valor de centralidade, observa-se um incremento na capacidade de propagação de informação de toda a rede. A média da capacidade de propagação na rede adaptada é maior que a capacidade de propagação do vértice mais influente da rede sem adaptação.

1.5 Estrutura do documento

Este trabalho está organizado da seguinte maneira:

- No Capítulo 2 são apresentados os conceitos relacionados ao trabalho como: conceitos básicos de redes complexas, modelos de redes artificiais considerados, uma descrição das medidas centrais em escala local e global adotadas no trabalho. Além disso, é apresentada uma revisão atual sobre dinâmicas de propagação baseadas em redes, especificamente, considerando os processos de difusão de epidemias e rumores.
- No Capítulo 3 são apresentadas as abordagens gerais adotadas neste projeto, como o modelo de propagação de informação adotado, a definição e configuração geral das simulações e as redes complexas consideradas no projeto.

- No Capítulo 4 é apresentado o modelo proposto de propagação heterogênea de rumores junto com a motivação da proposta, trabalhos relacionados, a metodologia e os resultados em redes artificiais e reais. Além disso, apresentamos os resultados de dois cenários ou casos de estudo, de melhoria ou controle da propagação, para exemplificar o alcance do modelo.
- No Capítulo 5 é apresentada a proposta do método de seleção de vértices influentes baseado nas comunidades da rede. A motivação do problema, os trabalhos relacionados e a metodologia são também apresentados. Adicionalmente, os resultados relacionados a escolha do comportamento dos propagadores e a fração final de vértices informados a partir das multiplex sementes iniciais são analisados.
- No Capítulo 6 são apresentados os resultados da análises do impacto que a assortatividade da rede tem na propagação de informação. A motivação e a definição do problema de maximização de informação, a metodologia e os resultados em redes artificiais, reais e a distribuição das sementes são discutidas. Além disso, são apresentadas as considerações em relação a uma análise estatísticas dos resultados obtidos.
- No Capítulo 7 é apresentado o problema de identificação dos propagadores menos influentes da rede. A motivação do problema, a definição dos vértices menos influentes e a metodologia de identificação mediante análise topológica ou supervisionada são descritas. Também, é apresentada a proposta e resultados de um método adaptativo para aumentar a capacidade de propagação dos vértices menos influentes e de toda a rede.
- Finalmente, no Capítulo 8 são apresentadas as contribuições e publicações geradas, as limitações e trabalhos futuros.

FUNDAMENTAÇÃO TEÓRICA

As redes sociais são a representação de um sistema complexo definido pelo conjunto de dois elementos: atores (pessoas, instituições ou grupos, que são os vértices da rede) e suas conexões (interações, relações, comunicação ou laços sociais) (WASSERMAN; FAUST, 1995). Ao analisar as redes sociais podemos descrever e medir as relações entre os atores (tangíveis ou não) e suas propriedades topológicas. Por exemplo, podem-se definir os vértices mais importantes ou centrais de uma comunidade, tanto pela importância topológica, quanto na sua capacidade de propagação de informação. Pesquisas e aplicações de impacto em redes sociais têm sido reportadas em diferentes contextos relevantes na sociedade. Casos como terrorismo (KREBS, 2002), prevenção e detecção de fraudes (WORRELL; WASKO; JOHNSTON, 2011), epidemias, como H1N1 (AMARAL L.A.N, 2004; KADUSHIN, 2005; HETHCOTE, 2006; BAJARDI *et al.*, 2011) ou de rumores (KEMPE; KLEINBERG; TARDOS, 2003), criminologia e economia (KADUSHIN, 2005), entre outros, são exemplos da relevância do tema. Desse modo, modelos de propagação de informação que utilizam o ferramental da Teoria de Redes Complexas, auxiliam no entendimento de como as informações se propagam na rede, como propor técnicas adequadas para controlar a difusão e como a estrutura da rede afeta a dinâmica, entre muitas outras.

2.1 Redes complexas

Vivemos rodeados de vários tipos de sistemas complexos, desde as células, constituídas por milhões de moléculas, até a sociedade, que é uma coleção de milhões de indivíduos e suas interações. Esses sistemas podem ser representados no formato de uma rede. Desse modo, existe todo tipo de redes e em toda escala. O cérebro é uma rede de neurônios conectados por axônios, enquanto que uma simples célula é uma rede de moléculas conectadas por reações bioquímicas. A sociedade, é uma rede de pessoas conectadas por relacionamentos de amizade, família, colegas, conhecidos, entre outros. Em maior escala, ecossistemas inteiros representam redes de espécies e

de cadeias alimentar. Além do mais, existem também várias redes tecnológicas, como a Internet, as redes de transmissão de energia elétrica, de transporte, entre muitas outras (NEWMAN, 2003; COSTA *et al.*, 2007). Inclusive, até a linguagem pode ser considerada como uma rede de palavras conectadas por valores semânticos da própria língua. São várias as ciências que participam no estudo dos sistemas complexos, como a teoria dos grafos, sistemas dinâmicos, mecânica estatística, teoria de jogos, processos estocásticos, probabilidade, entre outros. Além disso, o estudo de sistemas complexos abrange diferentes áreas como a biologia, a medicina, a economia, a sociologia, a computação, a física e engenharia.

Desse modo, o estudo de redes complexas permite o entendimento de muitos sistemas complexos, a partir de sua estrutura e da influência que esta tem sobre o comportamento da rede. As redes complexas possuem um grande número de elementos e conexões, além de ser dinâmicas alterando a sua estrutura ao longo do tempo. O interesse da pesquisa em redes está em entender a estrutura, suas propriedades e como a organização da rede pode influenciar no comportamento e evolução de processos dinâmicos, além da generalização para diferentes sistemas da natureza. Independente de função, idade ou escopo, as redes reais convergem a arquiteturas similares e apresentam propriedades que podem ser caracterizadas através do uso de medidas (BARABÁSI, 2009). Características como a distribuição de grau, a presença de *hubs*, a assortatividade nas conexões dos vértices, entre outras propriedades estruturais, são de grande relevância para o estudo da rede.

Neste capítulo são apresentados os conceitos relacionados às redes complexas e as medidas propostas para caracterizá-las. Os conceitos iniciais de redes complexas são apresentados na Seção 2.1.2; medidas de caracterização local são apresentados na Seção 2.1.3; na Seção 2.1.4 são apresentadas as medidas de escala global; Finalmente a Seção 2.1.2 apresenta alguns modelos gerais sobre tipos de redes complexas.

2.1.1 Conceitos básicos

Uma rede é um conjunto de elementos chamados vértices ou nós, os quais estão unidos entre si mediante conexões chamadas arestas. Formalmente definimos a rede $G = (V, E, W)$, sendo $V = \{v_1, v_2, \dots, v_N\}$ o conjunto de N vértices, $E = \{e_1, e_2, \dots, e_M\}$ o conjunto de M conexões ou arestas e $W = \{w_1, w_2, \dots, w_M\}$ os pesos das arestas que determinam a força da interação entre os respectivos vértices, no caso de redes ponderadas. Uma aresta pode ser direcionada, indicando o vértice inicial e o vértice destino da relação, ou pode ser não direcionada, onde ambos são o início e fim da relação. Em termos computacionais, esta estrutura pode ser representada por meio de uma matriz de adjacência A . Quando a rede é não direcionada, os vértices i e j estão ligados se a entrada $A_{ij} = A_{ji}$ na matriz é igual a 1. Caso contrário, será igual a 0. Neste trabalho são consideradas redes não direcionadas ou simétricas.

Dois vértices são chamados de vizinhos se estão conectados por uma aresta. O conjunto de vértices que se conectam com um vértice i é chamado de vizinhos de i ou vizinhança, $K(i)$.

O grau de um vértice, k_i , representa o tamanho de sua vizinhança ou o total de vizinhos que ele possui.

Para uma rede $G(N, M)$, o número de arestas M pode estar no intervalo $[0, M_{\max}]$, onde $M_{\max} = N(N - 1)/2$ é o número máximo de arestas da rede. A rede é chamada de esparsa se $M \ll N^2$ e densa se $M = O(N^2)$. Quando $M = M_{\max}$, a rede é chamada de rede completa G_N . Define-se que uma rede é regular, G_r , quando todos os vértices possuem o mesmo número r de conexões.

A caminhada de um vértice i até o vértice j é uma sequência consecutiva de arestas $(e_1, \dots, e_l)$ através de vértices adjacentes que começam em i e terminam em j . A distância caminhada é definida como o número l de arestas que estão na sequência. Uma trilha é uma sequência em que nenhuma aresta é repetida. Um caminho é uma caminhada em que nenhum vértice é visitado mais de uma vez. O caminho com menor distância entre dois vértices é conhecido como caminho mínimo ou caminho geodésico $(\ell_{i,j})$. Um ciclo é um caminho circular ou fechado, de pelo menos 3 vértices, onde nenhuma aresta é repetida começando e terminando no mesmo vértice. A rede G é chamada de conectada ou conexa se para qualquer par de vértices i e j distintos existe um caminho na rede que os comunica, caso contrário, é chamada de desconexa. A rede que não possui ciclos é chamada de acíclica ou de floresta. Uma floresta conexa é conhecida como uma árvore.

Os menores caminhos entre todos os vértices em uma rede podem ser representados através de uma matriz de distâncias $G\ell$, cujos elementos ℓ_{ij} expressam o valor do comprimento do caminho geodésico entre os vértices i e j . Estes caminhos podem ser obtidos por meio dos algoritmos de Dijkstra, BellmanFord ou busca em amplitude como o método Breadth-first (BRANDES, 2001). O valor $\ell_{\max} = \max_{i,j} \{\ell_{ij}\}$ é chamado diâmetro da rede.

Em redes simétricas, um componente é o maior sub-conjunto de vértices no qual existe ao menos um caminho entre cada par de vértices e nenhum caminho com algum vértice fora do conjunto. Caso i e j não pertençam ao mesmo componente, o conceito de distância não está bem definido, mas neste trabalho usamos como convenção $\ell_{ij} = \infty$. Logo, geralmente considera-se apenas o maior componente ou componente principal C_p para a caracterização da rede.

Existem outras medidas para caracterização topológica da rede. Algumas delas utilizam os conceitos de distância e caminhos mínimos. Outras analisam a distribuição de grau dos vértices ou a relação entre os vértices e seus vizinhos. Nas seguintes seções serão abordadas algumas medidas para redes não direcionadas, classificando-as entre medidas centrais ou locais e medidas globais ou da rede.

2.1.2 Modelos de redes artificiais

As redes também podem ser classificadas segundo a natureza dos dados que representam. Espera-se que redes de natureza iguais compartilhem propriedades semelhantes entre si. Newman

(2003) classifica as redes reais em quatro categorias: i) Redes sociais, ii) Redes de informação, iii) Redes tecnológicas e iv) Redes biológicas. As redes sociais são aquelas em que os vértices representam atores e as arestas as relações entre os atores para uma comunidade. Redes baseadas na troca de emails, ou redes como *Facebook* ou *Twitter*, formam parte desse conjunto.

As redes de informação são aquelas cuja topologia se assemelha com a estrutura da informação armazenada nos vértices. No caso, redes de citações de autores ou a *World Wide Web* (Web) são alguns exemplos. As redes tecnológicas são aquelas cujo propósito é a transmissão e distribuição de algum bem ou recurso. Exemplos são as redes de força elétrica e inclusive a própria Internet e sua estrutura física. Por último, as redes biológicas são todas aquelas que mapeiam algum sistema de seres vivos ou orgânico. Elas abrangem desde redes de proteínas e redes genéticas, considerando também redes corticóides e neuronais, até as redes que representam grandes ecossistemas.

Segundo o tipo de rede existem propriedades, como a distribuição de grau dos vértices, proporção de triângulos, média dos caminhos mínimos, assortatividade, entre outros, que a caracterizam. Os modelos de redes artificiais buscam adotar algumas dessas características para isolar e estudar por separado situações particulares. Desse modo, nos modelos são controlados alguns desses parâmetros e outros são desconsiderados. Quanto mais parâmetros forem controlados pelo modelo, maior complexidade apresentará. A seguir são descritos alguns modelos de redes artificiais que serão considerados neste trabalho.

2.1.2.1 Modelos de redes aleatórias

As redes aleatórias foram inicialmente propostas pelos matemáticos Erdős e Rényi (1959) e, desde então, muitos estudos teóricos sobre essas redes têm sido desenvolvidos (BARABÁSI; ALBERT, 1999; BOLLOBÁS, 2001). Os modelos fixam alguns parâmetros da rede, como o número de vértices, de arestas, ou uma probabilidade de conexão, de modo a obter redes com grau médio semelhante.

Existem dois modelos de redes aleatórias que se destacam e que são conhecidos pela notação matemática $G(N, M)$ e $G(N, P)$. No modelo $G(N, M)$, ou modelo ER (ERDÖS; RÉNYI, 1959), distribui-se sem nenhuma preferência M arestas entre N vértices de maneira uniforme, evitando a auto conexões e multi arestas. As redes possuem um número fixo de conexões, conhecido como *canonical ensemble*. No modelo $G(N, P)$, cada par de vértices é testado uma única vez, conectando-os com uma probabilidade P . O modelo gera redes com um número de conexões variável, conhecido como a propriedade *grand-canonical ensemble*.

Como podem ser geradas diferentes redes mediante os mesmos parâmetros, considera-se a esperança dos valores utilizando um conjunto (*ensemble*) de redes. A probabilidade de que uma realização do modelo $G(N, P)$ produza uma rede com M arestas é dada por P_M .

Como P_M segue uma distribuição binomial, pode-se estimar o número esperado de arestas

Figura 1 – Distribuição de grau de 30 iterações para uma rede ER com $N = 500$, $P = 0.045$, $\langle k \rangle = 23.76 \pm 0.33$, e $\langle \ell \rangle = 2.223 \pm 0.10$

Fonte: Elaborada pelo autor.

segundo $\langle m \rangle \sim \frac{N(N-1)}{2}P$. A probabilidade de que um vértice i tenha k arestas é dada por

$$P(k) = \binom{N-1}{k} P^k (1-P)^{N-1-k}, \quad (2.1)$$

na qual, P^k é a probabilidade de sortear as k arestas, $(1-P)^{N-1-k}$ é a probabilidade de que não existam as $(N-1-k)$ arestas restantes e o combinatório representa as diferentes possibilidades de selecionar os vértices finais das k arestas. Na rede aleatória, os vértices são estatisticamente equivalentes. Assim, a probabilidade de que um vértice escolhido tenha grau k segue a distribuição $P(k) \sim \text{Binomial}(N-1, P)$, uma vez que o processo de geração das redes é um processo de Bernoulli homogêneo. Desse modo, o número de conexões esperado é dado por $\langle k \rangle = (N-1)P$ (COSTA *et al.*, 2007). Considerando o tamanho da rede no limite, $N \rightarrow \infty$, e o grau médio constante, temos que a distribuição de grau aproxima-se a uma distribuição Poisson (BOLLOBÁS, 2001), isto é.

$$P(k) = \frac{e^{-\langle k \rangle} \langle k \rangle^k}{k!}. \quad (2.2)$$

Isso significa que a maioria dos vértices terá grau próximo de $\langle k \rangle$ e não existe a presença de *hubs* na rede, o qual pode ser comprovado com o segundo momento da rede. Além disso, a média dos caminhos mínimos é pequena para esse tipo de redes, crescendo proporcionalmente com o logaritmo do número de vértices, com $\langle \ell \rangle \sim \frac{\log(N)}{\log(\langle k \rangle)}$. Exemplo dos valores experimentais para redes ER são apresentados na Figura 1.

Os modelos apresentados não consideram outras propriedades na construção das redes. Por tal motivo, eles não são adequados para representar redes reais (BARABÁSI, 2001; NEWMAN, 2003). Na seguinte seção é apresentado um modelo que considera a proporção de triângulos na rede, propriedade observada em redes reais (NEWMAN, 2010).

Figura 2 – Modelo de Watts-Strogatz para redes mundo pequeno. Adaptada de Watts e Strogatz (1998a)

Fonte: Elaborada pelo autor.

2.1.2.2 Modelos de redes de pequeno mundo

Em muitas redes reais, apesar de serem grandes, há um caminho relativamente curto entre quaisquer pares de vértices. Essa característica é conhecida como propriedade de pequeno mundo (*small-world*) e foi inicialmente observada por Milgram (1967). Em um experimento enviando cartas nos Estados Unidos, Milgram encontrou que a distância média entre as pessoas em uma rede social é muito pequena, em média, seis passos eram suficientes para conectar duas pessoas quaisquer, dando origem a famosa hipótese dos seis graus de separação.

De igual maneira, Watts e Strogatz (1998a) notaram que as redes reais possuíam essa característica, além de uma grande presença de triângulos (ciclos de ordem três), muito maior em comparação com redes aleatórias semelhantes (com valores iguais de N e M). Desse modo, foi inicialmente evidenciado que as redes reais não são totalmente aleatórias, e que existe algumas regras que governam sua formação. Foi proposto um método de construção de redes conhecido como WS, no qual é possível variar a proporção de triângulos na rede. O modelo consiste em um processo de re-ligação de arestas dada uma probabilidade P .

O processo começa com um lattice unidimensional regular de N vértices, no qual cada vértice é simetricamente ligado aos seus $2k$ vizinhos mais próximos para um total de $M = kN$ arestas. Para cada vértice, cada aresta ligada a um vizinho no sentido horário é re-ligada a outro vértice escolhido aleatoriamente com uma probabilidade P , ou preservada com uma probabilidade $1 - P$. Nota-se que para $P = 1$ o modelo gera uma rede aleatória, com baixa quantidade de ciclos de ordem três, com caminho médio pequeno mas com a restrição de que cada vértice tem conectividade mínima $k_{min} = k$. A rede aleatória é diferente do modelo ER, pois o número mínimo de arestas é mantido mas o grau dos vértices é diferente. Quando $P = 0$, o modelo gera redes com alta presença de triângulos e regular.

Na Figura 2 observamos o modelo WS proposto por Watts e Strogatz, com apenas um

Figura 3 – Característica do modelo WS para rede com $N = 600$: (a) e (b), distribuição de grau segundo P ; (c) média do coeficiente de *clustering* $\langle C \rangle$ VS a média dos menores caminhos geodésicos $\langle g \rangle$ com $m = 6$ para 30 iterações

Fonte: Elaborada pelo autor.

componente. Diferente do modelo *ER*, a medida que P aumenta, o número de recombinação de arestas também aumenta, interpolando de uma rede regular até uma rede aleatória. No meio termo são encontradas as redes de pequeno mundo do modelo (Figura 3)

Na rede regular circular descrita anteriormente, cada vértice tem o mesmo grau m (Figura 3a). Uma vez feitas as recombinações entre as arestas, a rede começa se tornar em pequeno mundo. A definição para a distribuição de grau do modelo *WS* é apresentado a seguir

$$P_P(k) = \sum_{n=0}^{\min(k-m, m)} \left[\binom{m}{n} (1-P)^n P^{m-n} \right] \left[\frac{(mP)^{k-m-n}}{(k-m-n)!} e^{-Pm} \right], \quad (2.3)$$

onde o primeiro termo representa a probabilidade de que sejam mantidas as conexões e o segundo termo a distribuição de probabilidade das conexões que são recombinadas, dado um N grande. No caso em que $P \rightarrow 1$ temos que $P_P(k)$ segue uma distribuição de Poisson (Figura 3b), com $k' = k - m$ e $\langle k' \rangle = m$, isto é

$$P(k) = \frac{(m)^{k-m}}{(k-m)!} e^{-m}. \quad (2.4)$$

Considerando $P = 0$, o número esperado de triângulos ou conexões entre vizinhos para

cada vértice é $C_i = 3m(m-1)/2$ e o número total esperado de conexões seria $C_t = 2m(2m-1)/2$. Assim, pode-se definir o valor máximo do coeficiente de agrupamento fazendo C_i/C_t , obtendo como resultado $C_{\max} = \frac{3(m-1)}{2(2m-1)}$, que para fins práticos e valores grandes de m , temos que $C_{\max} \leq 3/4$.

Na Figura 3c temos a relação do coeficiente de *clustering* e a média dos caminhos geodésicos da rede em função de P , calculado para 30 simulações. A linha pontilhada apresenta o comportamento de $\langle C \rangle$, com seu máximo teórico $C_{\max} = 0.6$ quando $P = 0$. A linha contínua apresenta o comportamento da média dos caminhos geodésicos da rede, também como uma fração de seu valor máximo $g_{\max} = n/2m = 50$. Observa-se que existe um intervalo de valores de P nos quais o modelo apresenta de forma simultânea os efeitos de um coeficiente de *clustering* alto e média de caminhos mínimos significativamente baixos, causando o efeito de pequeno mundo.

2.1.2.3 Modelos de redes sem escala

Nos modelos anteriores, a probabilidade de dois vértices se conectarem está definida por uma probabilidade constante P , seguindo uma distribuição Poisson para o grau dos vértices. Em contraste, Barabási e Albert (1999) encontraram que em muitas redes reais existe um número abundante de vértices com poucas conexões e uma pequena minoria de vértices que apresentam a maior parte das conexões. Esse tipo de rede segue uma distribuição lei de potência (ALBERT; BARABÁSI, 2002; NEWMAN, 2003; CLAUSET; SHALIZI; NEWMAN, 2009; BARABÁSI, 2009), do tipo $P(k) \sim k^{-\gamma}$. A distribuição de grau é altamente heterogênea, com alguns poucos *hubs* ligados a muitos vértices que possuem graus baixos.

As redes sem escala são um subconjunto das redes com distribuição de lei de potência, dado que o fator de potência γ varia entre $[1, 3]$. São denominadas sem escala porque a média do grau $\langle k \rangle$ não representa um valor significativo de escala na rede. Entre as principais características das redes sem escala observa-se:

- A maioria dos vértices têm grau baixo e só poucos vértices têm grau alto (presença de *hubs*);
- São redes dinâmicas, que podem crescer à medida que passa o tempo.
- São redes resistentes a falhas ou erros. Uma falha é considerada como a remoção uniforme de vértices da rede. No caso, a rede continuará conectada e não se fragmentará.
- Exibem uma média de caminhos mínimos pequena, tendo assim o fenômeno de mundo pequeno.

Barabási e Albert (1999) propõem um modelo que gera redes sem escala com $\gamma = 3$, conhecido como redes BA. O processo inicia com m_0 vértices totalmente conectados. A cada

Figura 4 – Modelo de rede sem escala Barabási e Albert: (a) exemplo de uma rede BA com $N = 500$, $m_0 = m = 3$ e $\gamma = 2$; (b) distribuição de grau para 30 simulações com os mesmos parâmetros e resultado de $\langle k \rangle = 5.976 \pm 0.0$, $\langle \ell \rangle = 2.749 \pm 0.048$ e $\langle C \rangle = 0.215 \pm 0.026$

Fonte: Elaborada pelo autor.

passo de tempo $t = 1, 2, 3 \dots N - m_0$ um novo vértice j com m arestas é adicionado à rede. A probabilidade que o novo vértice j se conecte com um vértice da rede é dada pela Equação 2.5.

$$\Pi(k_i) = \frac{k_i}{\sum k_j}, \quad (2.5)$$

onde o denominador é a soma do grau para todos os vértices. Esta abordagem faz com que os vértices com maior número de conexões tenham maior probabilidade de se conectar com os novos vértices. Pela regra de conexão preferencial, as conexões entre os vértices não acontecem de modo randômico e sim de modo preferencial, onde os “ricos ficam cada vez mais ricos”. Todos os vértices novos com suas m arestas são conectados seguindo esse procedimento.

As seguintes expressões analíticas podem ser obtidas a partir do modelo (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008): o grau médio da rede pode ser calculado segundo $\langle k \rangle \sim 2m$; a média dos caminhos mínimos $\langle \ell \rangle \sim \log(n)/\log(\log(n))$, o qual é conhecido como a propriedade *ultra small-world*; no caso da média do coeficiente de *clustering*, conclue-se que $\langle C \rangle \sim \frac{M}{8N^2} \log(N)^2$. Um exemplo de rede BA com a distribuição de grau e os valores das medidas aparece na Figura 4. Observa-se que a variância da distribuição é muito maior por causa da presença dos *hubs* na rede, em comparação aos modelos ER e WS. Essa condição é mais clara e evidente na seguinte seção, onde o modelo proposto interpola de redes sem escala até redes randômicas.

2.1.2.4 Modelo de rede sem escala espacial

Apresentamos um modelo que produz redes sem escala que possuem característica espacial. Redes espaciais são aqueles cujos vértices estão também distribuídos no espaço bidimensional, sendo também grafos planares. Barthélemy (2003) propôs um modelo de rede onde

os vértices estão uniformemente localizados no espaço considerando uma *lattice* d dimensional regular com comprimento l , simulando uma rede dispersa geograficamente. O modelo adota os conceitos de seleção de proximidade entre os vértices, sem importar o *kernel* para o cálculo da distância, junto com a regra de conexão preferencial. Portanto, a probabilidade de conectar um vértice entrante x é proporcional a

$$\prod(x, y) \propto \frac{k_y + 1}{\exp(d_{xy}/r_c)}, \quad (2.6)$$

onde d_{xy} é a distância euclidiana entre o vértice x e y , k_y é o número de conexões de y , e r_c é um parâmetro de escala finita que controla a correlação de grau-grau (NEWMAN, 2003) e a proporção de triângulos (WATTS; STROGATZ, 1998a) da rede gerada.

O modelo de rede espacial sem escala (SSF) (BARTHÉLEMY, 2003), do inglês *spatial scale-free model*, inicia com todos os vértices distribuídos no espaço e executando os seguintes passos: (i) seleciona aleatoriamente n_0 vértices ativos iniciais; (ii) escolhe um vértice inativo x aleatoriamente e o conecta com algum ativo segundo a probabilidade $\prod(x, y)$, tornando o vértice x em ativo; (iii) os dois últimos passos são repetidos até que todos os vértices sejam ativados.

O grau médio da rede é obtido ao repetir m vezes o segundo e terceiro passos para cada vértice, obtendo-se como resultado $\langle k \rangle = 2m$. O modelo produz redes planares inseridas no espaço com distribuição de grau seguindo uma lei de potência. Para a geração das redes, nós consideramos aqui os valores dos parâmetros $r_c = 0,05$, $L = 1$ e $d = 2$, similares aos empregados no trabalho original (BARTHÉLEMY, 2003).

2.1.3 Medidas de centralidade

As medidas de centralidade fazem referência a que tão importante ou central é um vértice na rede de acordo com alguma propriedade. Tais medidas geralmente são usadas para analisar a influência de vértices em processos dinâmicos. Por exemplo, é possível prever o comportamento da rede segundo as condições que governam os vértices, a estrutura e o contexto do fenômeno mapeado. Ou também, o processo de propagação de um vírus ou de uma informação em uma rede depende da topologia da rede e das relações de seus vértices. Do mesmo modo, a estratégia para controlar a propagação de um vírus em uma rede tecnológica pode ser diferente do método de controle de uma epidemia em uma sociedade (BARABÁSI, 2007). Atualmente, não existe um método universal que identifique qual é a melhor medida central para prever o comportamento de um processo dinâmico, pois existem muitas propostas e a grande maioria estão correlacionadas entre si (COSTA *et al.*, 2007).

As medidas de centralidade buscam identificar o quanto um vértice é importante em um determinado escopo. Dentre alguns desses escopos temos, a definição de popularidade de um vértice (centralidade de grau), o que tão próximo é um vértice dos demais (centralidade de proximidade) e o que tão concorrido pode ser um vértice na transmissão de informação

na rede (centralidade de intermediação), ou que tão central está um vértice do núcleo da rede (centralidade de núcleo). Por exemplo, podemos pensar no problema da localização de um hospital ou de um centro comercial. No caso do hospital, é interessante encontrar um local que esteja o mais próximo de todos os pontos da cidade. Por outro lado, o centro comercial pode ser construído próximo ao ponto onde há mais tráfego de pessoas.

2.1.3.1 Centralidade de grau (DG)

O grau ou conectividade de um vértice i , denominado k_i , é o número de arestas ou conexões que incidem sobre ele. O grau do vértice i é definido em termos da matriz de adjacência como

$$k_i = \sum_{j \in N} A_{ij}. \quad (2.7)$$

Em termos de redes sociais, um indivíduo com um grau alto possui muitos amigos ou conexões e portanto, pode ser importante. Por exemplo, atores famosos são seguidos nas redes sociais por muitas pessoas no mundo, páginas Web importantes são bastante referenciadas, pesquisas destacadas são aquelas que possuem um maior número de citações. Desse modo, vértices centrais ou importantes na rede podem ser aqueles que possuem um número de conexões elevado em comparação com os demais.

O grau médio $\langle k \rangle$ representa a esperança da distribuição do grau para os vértices da rede, e define-se como $\langle k \rangle = \frac{1}{N} \sum_{i \in N} k_i = 2M/N$. No caso de redes direcionadas, considera-se $\langle k \rangle = \langle k^{in} \rangle + \langle k^{out} \rangle$. Os vértices que têm um grau elevado são chamados de *hubs*. Na Figura 5a temos a centralidade de grau para os vértices de uma rede real. Quanto mais vermelho escuro o vértice, maior seu grau.

Outra medida relacionada com conectividade é a média do grau dos vizinhos $knn(i)$ (COSTA *et al.*, 2007). O cálculo do grau médio dos vizinhos para o vértice i é definido como,

$$Knn(i) = \frac{1}{k_i} \sum_{j=1}^N k_j A_{ij}, \quad (2.8)$$

onde k_i é o total de vizinhos de i , k_j é o total de vizinhos de j e a matriz de adjacência A determina a conexão entre i e j . Essa medida indica quantos contatos na média seus vizinhos tem.

2.1.3.2 Grau de intermediação (BE)

Para quantificar o tráfego de informação que passa por um vértice i é utilizada a medida de grau ou *BE*. Dado um vértice i , calcula-se o número de caminhos mínimos que conectam

cada par de vértices não adjacentes (j, k) que passam por i , sobre o total dos caminhos mínimos entre (j, k) . Assim, o vetor de *betweenness centrality* do vértice i é dado por

$$B_i = \sum_{j,k \in V, j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}, \quad (2.9)$$

onde σ_{jk} é o total dos caminhos mínimos entre j e k , e $\sigma_{jk}(i)$ é o número de caminhos mínimos entre j e k que passam por i . Com esta medida é possível saber o quanto um vértice é essencial para o tráfego na rede. O conceito de *betweenness centrality* também pode ser estendido para as arestas. O *betweenness centrality* de uma aresta se define como o número de caminhos mínimos entre pares de vértice que passam por essa aresta.

O cálculo do grau de intermediação pode ser generalizado como uma medida da rede inteira. Para isso, se calcula a média dos valores de *betweenness centrality* de todos os vértices da rede, como definido por

$$\langle B \rangle = \sum_i^N \frac{B_i}{N}. \quad (2.10)$$

Observa-se que o valor máximo de *betweenness centrality*, $B_{\max} = \max_i \{B_i\}$, é diferente para redes de tamanho diferente. Logo, não é possível comparar $\langle B \rangle$ entre diferentes redes. No entanto, os valores de B_i podem ser normalizados dividindo-os pelo maior valor possível de *betweenness centrality* da rede $B_{\max} = (N-1)(N-2)/2$. Desse modo, todos os valores estarão entre 0 e 1. Na Figura 5a temos como exemplo o nível de centralidade *betweenness* para uma rede real. Observa-se que não necessariamente os vértices com maior grau são aqueles com maior *betweenness* na rede, como no caso do vértice 37. Também, observamos que os vértices com maior grau de intermediação servem como ponte de transmissão entre os dois grupos.

A partir do grau de intermediação é possível definir medidas que ajudem a caracterizar a estrutura da rede. Freeman (1977) propõe a medida do ponto central de dominância, que é dada por,

$$CPD = \frac{1}{N-1} \sum_i (B_{\max} - B_i), \quad (2.11)$$

na qual, $B_{\max}$, como foi explicado anteriormente, é o maior valor de intermediação possível da rede. O valor dessa medida será 0 ao se tratar de uma rede completa e terá valor de 1 quando se tratar de uma rede em formato de estrela, com um único vértice central e considerando que os valores de intermediação estão normalizados.

2.1.3.3 Medidas de vizinhança

A transitividade é uma propriedade visível nas redes reais. Em termos sociais, significa que se $\mathcal{A}$ é amigo de $\mathcal{B}$ e $\mathcal{B}$ é amigo de $\mathcal{C}$, existe uma forte tendência de $\mathcal{A}$ ser amigo de $\mathcal{C}$.

Figura 5 – Análise de centralidade para a rede Dolphins (LUSSEAU *et al.*, 2003): (a) quanto mais vermelho escuro maior o grau e quanto maior o vértice maior seu grau de intermediação; (b) quanto mais azul escuro menor *eccentricity* e quanto maior o vértice maior o valor do *PageRank*

(a) Grau e grau de intermediação

(b) Excentricidade e *PageRank*

Fonte: Elaborada pelo autor.

Topologicamente significa a presença de ciclos ou caminhos fechados de ordem três na rede. O coeficiente de transitividade da rede proposto por [Barrat e Weigt M. \(2000\)](#), define-se por

$$C_t = \frac{3 \times \text{Num de triângulo na rede}}{\text{Num vértices conectados às triplas}}. \quad (2.12)$$

O fator de 3 no numerador indica que cada triângulo contribui com 3 triplas. Em termos práticos, C_t é a probabilidade de que dois vértices que são vizinhos de outro vértice sejam também vizinhos entre si.

Além do cálculo da transitividade para a rede toda (Equação 2.12), existe outra abordagem proposta denominada CT . O coeficiente de aglomeração é definido localmente para cada vértice, por

$$C_i = \frac{\text{Num de triângulos conectados ao vértice } i}{\text{Num de triplas centradas no vértice } i} = \frac{2e_i}{k_i(k_i - 1)}. \quad (2.13)$$

No caso em que o grau do vértice i seja 0 ou 1, o valor de C_i é definido como 0. C_i terá valor 1 se todos seus vizinhos estiverem conectados entre si.

O coeficiente de aglomeração também pode ser usado para caracterizar a rede de forma global. Define-se $\langle CT \rangle$ como a média dos valores de C_i para todos os vértices da rede.

$$\langle CT \rangle = \frac{1}{N} \sum_{i=1}^N C_i. \quad (2.14)$$

Nota-se que as duas abordagens de coeficiente de aglomeração global produzem resultados diferentes. A Equação (2.14) dá mais peso as contribuições dos vértices com menor grau. Também, temos que a primeira abordagem é fácil de ser calculada analiticamente, porém, a segunda abordagem é muito mais fácil de ser calculada computacionalmente.

2.1.3.4 Centralidade de proximidade (CL)

CL

A proximidade ou *closeness* é definida como a média das distâncias entre todos os vértices da rede ([SABIDUSSI, 1996](#)). Por exemplo, um vértice pode estar distante do grupo que pertence e ainda ter um grau alto ao estar bem conectado em escala global. O grau de proximidade Cl_i é o inverso da média das distâncias geodésicas de todos os nós para o i -ésimo nó da rede, Isto é,

$$Cl_i = \ell_i^{-1} = \frac{N}{\sum_{j \neq i} \ell_{ij}}, \quad (2.15)$$

onde ℓ_{ij} é a distância geodésica do vértice i até o vértice j . Essa medida explora a relação de um vértice com todos os demais. Quanto maior o valor de Cl_i , menor a distância desse vértice com os outros indivíduos da rede. Como pode ser observado, a medida de *closeness* considera em média que tão distante está um vértice dos demais. Por exemplo, podemos utilizar essa medida

para encontrar um ponto em uma cidade que seja apropriado para uma biblioteca ou hospital. O ponto da cidade com maior Cl_i indicara que em média está mais próximo de todos os pontos da cidade.

Também, existe outra medida interessante para caracterizar posições estratégicas dos vértices relacionada com a distância na rede. A medida *eccentricity* ou excentricidade avalia o quão central ou próximo está um vértice de aquele mais afastado a ele. Para o vértice i , define-se como

$$ec(i) = \max_{i \neq j} \{\ell_{ij}\}, \quad (2.16)$$

no qual, quanto menor o valor mais central o vértice. O raio da rede define-se como a menor excentricidade entre os vértices, com $rad_G = \min_i \{ec(i)\}$. Na Figura 5b temos os valores de excentricidade para uma rede real. Observa-se que os vértices com os menores valores da medida estão localizados no centro da rede. Isso implica que ao contrário do exemplo da biblioteca ou do centro comercial, alocar um hospital exige que seja fácil e rápido chegar a ele desde qualquer ponto da cidade. Desse modo, os vértices com melhor excentricidade podem ser mais apropriados por estarem no centro da cidade.

2.1.3.5 Centralidade de autovetor (EV)

Quando definimos o grau como medida de centralidade, consideramos cada vizinho com o mesmo peso. No entanto, nem todos os vizinhos são equivalentes, a menos que a rede seja regular. Esta medida pode ser considerada como o grau de importância de um vértice relacionado com a importância de suas conexões ou vizinhos. Em muitas situações, um vértice é importante em uma rede se ele tem muitas conexões ou se ele está conectado com vértices importantes. A centralidade de autovetor foi proposta por Bonacich (1972), que argumentou que o autovetor associado ao maior autovalor da matriz de adjacência pode ser um bom descritor de centralidade, matematicamente

$$A\vec{X} = \lambda\vec{X}, \quad (2.17)$$

onde $\vec{X} = \{X_1, X_2, \dots, X_N\}$ é o vetor de centralidade e λ é o maior autovalor de A . Dado que A é simétrica e todo $A_{ij} \geq 0$, então pelo teorema de Perron-Frobenius temos que o maior autovalor é único e os elementos do autovetor correspondente são positivos. Computacionalmente, podemos definir o valor para cada X_i como um valor calculado a partir de um número de iterações, isto é

$$X_i^t = \sum_j A_{ij} X_j^{t-1}, \quad (2.18)$$

onde t é o passo das iterações e por definição para $t = 0$ temos $\vec{X}^0 = \{1, \dots, 1\}$. Foi adotado um tempo de iteração máximo $t_{\max} = 1000$ para todas as simulações apresentadas nesse

trabalho.

2.1.3.6 Centralidade PageRank (PR)

PageRank é a medida utilizada pela companhia *Google* para ranquear as páginas da Web de acordo com a sua importância. Introduzida por [Brin e Page \(1998\)](#) é o coração do motor de busca da companhia. A ideia básica é que se um vértice i recebe mais conexões do que um vértice j , então i deve ser mais importante do que j . Além disso, um vértice importante recebe conexões de outros vértices importantes.

Por definição, o algoritmo tenta simular o comportamento de um usuário enquanto está na Internet. Na maioria do tempo, o usuário visita páginas web enquanto navega, acessando *hiperlinks* da página atual. Por outro lado, o usuário pode pular para outra página web escrevendo uma nova URL no navegador, ou escolhendo um link favorito, entre outras possibilidades. A formulação da medida é apresentada na Equação 2.19

$$\pi_i^t = \alpha \sum_j \frac{A_{ij} \pi_j^{t-1}}{k_j^{out}}, \quad (2.19)$$

onde k_j^{out} é o grau de saída do vértice j , o qual dilui a importância de i . No caso de $t = 0$ temos que $\vec{\pi}^0 = [1/N, \dots, 1/N]_{1 \times N}$ é o vetor com a mesma probabilidade de escolha para qualquer vértice. Contudo, no caso em que $k_j^{out} = 0$ ou exista a presença de ciclos na rede, o valor $\vec{\pi}$ diverge. Para resolver essa questão é utilizado um ajuste estocástico, fazendo-se com que o vértice j tenha a mesma probabilidade de se conectar com todos os vértices da rede, definido $\tilde{k}_j^{out} = k_j^{out} + \delta(k_j^{out}, 0)(1/N)$. Também, pode ser utilizado um ajuste ergótico, fazendo com que todo i possa pular para j com uma probabilidade $(1 - \alpha)$. Por essa razão, na Equação 2.19 aparece o parâmetro α , que indica a probabilidade de navegar em sequência.

Neste trabalho foi adotado o valor de $\alpha = 0.85$ como definido por [Brin e Page \(1998\)](#). O valor de convergência adotado foi de $\varepsilon = 0.0010$. Na Figura 5b temos os valores de *PageRank* para uma rede social real. Observa-se que os vértices mais importantes não são necessariamente os que possuem maior grau, os que estão em média mais próximos de todos os vértices ou os que intermediam a comunicação entre duas comunidades. No exemplo podemos ver que os vértices com maior *PageRank* estão conectados de modo mais influente na rede.

2.1.3.7 Centralidade K-core (KC)

Um modo de descrever a topologia de uma rede real em termos de subredes é mediante o uso de uma decomposição via *k-core*. Uma subrede H_k é uma *k-core* ou núcleo de ordem k se para todo vértice i na subrede o grau $k_i \geq k$, e H_k for a maior rede com esta propriedade ([SEIDMAN, 1983](#)). O *core* de maior ordem é chamado de core principal de um vértice. Tem $Kc(i) = k$ se ele pertence a um *k-core* e não ao $(k-core + 1)$.

Figura 6 – Exemplo de decomposição em núcleos de uma rede. (a) representa o conjunto de vértices mais centrais com $Kc = 3$; (b) vértices no núcleo $Kc = 2$; (c) vértices no núcleo $Kc = 1$.

Fonte: Adaptada de Kitsak *et al.* (2010).

Os valores de k -core são obtidos ao remover da rede todos os vértices com grau menor ou igual que k . Depois dessa remoção, alguns vértices terão grau menor ou igual a k ou novos componentes se desconectam do componente principal. Esses vértices são removidos e a rede é analisada novamente. Quando nenhum outro vértice possa ser removido, a subrede não vazia resultante H_k é o núcleo de ordem k da rede original.

O Algoritmo 1 é uma proposta de Batagelj e Zaversnik (2003) com ordem de complexidade $O(M)$. A Figura 6 exemplifica o conceito de decomposição em núcleos para uma rede. Resalta-se que o processo de associar um valor de k -core é único para cada vértice. Assim, a medida tem sido utilizada na análise de percolação e propagação de epidemias e em análises de interações e predição de funções de proteínas (COSTA *et al.*, 2011).

2.1.3.8 Centralidade de buracos estruturais (HO)

Alguns vértices na rede funcionam como pontes entre grupos ou outros vértices e, se forem removidos, ocorreria um furo na estrutura. Os vértices com maior centralidade de buraco estrutural agem como elos entre grupos de vértices sem conexões diretas, sendo eles importantes para a conectividade de regiões locais. O fator analisado pela medida é a redundância na transmissão que cada vértice tem em sua vizinhança, avaliando se a sua localização e conexões trazem oportunidades no fluxo de transmissão de informação (BURT, 1992).

O algoritmo considera todos os vértices como redes do tipo ego, ou seja, redes onde as conexões não relacionadas o dito vértice não têm um efeito direto e são desconsideradas. Para cada vértice, o valor da medida é a fração de buracos isolados que surgem na sua rede de ego

Algoritmo 1 – Algoritmo de decomposição *k-core*

```

1: criar lista de subredes  $H$ 
2: gerar lista de grau e vértices  $K$ 
3: criar variável  $kcore = 1$ 
4: repita:
5:   repita:
6:     se  $Kc_i == kcore$ 
7:        $H_i = kcore$ 
8:       remover vértice  $i$ 
9:       remover sub-componentes
10:    atualizar  $K$ 
11:   fim se
12: enquanto  $\exists i | Kc_i \leq kcore$ 
13:   incremente  $kcore$ 
14: enquanto tamanho  $K > 0$ 
15: retornar  $H$ 

```

quando o vértice for removido. Quanto maior a fração de buracos estruturais associados ao vértice, maior sua centralidade. Sendo assim, vértices com alto grau de conectividade tendem a ter baixos valores de HO , uma vez que suas redes ego são maiores e mais densamente interconectadas, o qual diminui a fração de buracos isolados. A complexidade no tempo para calcular a medida para todos os vértices da rede é $\mathcal{O}(n + n * \langle k \rangle^2)$.

2.1.4 Medidas globais

Entender a topologia das redes e as propriedades emergentes que a caracterizam mediante uso de medidas apropriadas é um problema destacado na área de redes (NEWMAN, 2003). A seguir são apresentadas algumas medidas em escala global à rede.

2.1.4.1 Medidas relacionadas a conectividade

A distribuição do grau de uma rede representa a probabilidade de que um dado vértice escolhido de maneira aleatória (uniforme) tenha grau k . Um método para calcular a probabilidade de k é dado pela fórmula $p(k) = N_k/N$, onde N_k é número de vértices que tem grau k na rede. Contudo, se só expressamos os valores de $p(k)$ segundo o número de ocorrências de vértices com grau k , estamos adotando probabilidades de $p(k) = 0$ no caso em que nenhum vértice tenha grau k . Sendo assim, utiliza-se a função auxiliar do complemento da distribuição acumulada (Equação 2.20)

$$p_k = \sum_{k'=k}^{\infty} p(k') = p(K > k), \quad (2.20)$$

onde, p_k é a função de probabilidade para os vértices que tenham grau maior ou igual do que K . Essa técnica tem a vantagem de que a forma da distribuição é preservada, quando a distribuição é livre de escala. Ou seja, no caso em que a distribuição de grau seja lei de potência, a distribuição

acumulada também será uma lei de potência. Contudo, este método tem como desvantagem que ao ser acumulada, os pontos sucessivos são correlacionados e, portanto, não podem ser utilizados métodos tradicionais para estimar o valor de γ quando $P(k) = k^{-\gamma}$. Uma alternativa para isso é utilizar algum método estimador de variáveis, como por exemplo, a função de máxima verossimilhança sobre o conjunto de dados.

Podemos calcular os momentos da distribuição de grau, de acordo com

$$\langle k^n \rangle = \sum_{k=0}^{\infty} k^n p(k), \quad (2.21)$$

onde o segundo momento resulta em $\langle k^2 \rangle = \sum_{k=0}^{\infty} k^2 p(k)$.

Considerando a distribuição de grau junto com um fator de escala λ no qual $k \rightarrow \lambda k$, a distribuição continuará tendo a mesma forma, chamadas de redes sem escala. Isto significa que é possível caracterizar o nível de heterogenidade ou complexidade da rede, segundo

$$C = \frac{\langle k^2 \rangle}{\langle k \rangle}, \quad (2.22)$$

considerando o segundo e o primeiro momento da distribuição de grau (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008). De fato, as flutuações na complexidade são marcadas pela normalização da variância, a qual caracteriza as redes sem escala com $\langle k^2 \rangle \rightarrow \infty$, e no caso das redes homogêneas com $k_i \sim \langle k \rangle$.

Também, é possível calcular o nível de desordem das conexões entre os vértices. Com a entropia da distribuição do grau obtém-se o nível médio de heterogenidade da rede. Para isso, utiliza-se a versão normalizada da entropia de Shannon, isto é

$$\tilde{H} = -\frac{\sum_{k=0}^{\infty} p(k) \log(p(k))}{\log(N)} \quad \text{com } 0 \leq \tilde{H} \leq 1. \quad (2.23)$$

O máximo valor para a entropia acontece quando a distribuição de grau apresenta uma distribuição uniforme e o menor valor possível acontece quando todos os vértices têm o mesmo grau. A entropia de uma rede está relacionada com a robustez ou seu nível de resiliência (WANG *et al.*, 2006).

A robustez da rede também está relacionada com o correlação do grau da rede. Uma rede é assortativa (AN) se os vértices tendem a se conectar com vértices de grau semelhante. Em redes com assortatividade negativa ou disassortativas (DN), os vértices de grau baixo tendem a se conectar com vértices fortemente conectados (*hubs*). Pode acontecer que a rede não apresente nenhum dos casos anteriores, fazendo com que as conexões não tenham um padrão em comum. Esse tipo de rede é chamada de redes não assortativas ou não correlacionadas (NA). Para saber se uma rede é assortativa pode-se utilizar a média do grau dos vizinhos encontrando de forma gráfica a solução ou fazendo-se o cálculo do coeficiente de Pearson. Para a primeira abordagem, calcula-se a distribuição dos valores de $Knn(i)$ (Equação 2.8) da rede, que representa o $Knn(k)$,

definido

$$Knn(k) = \frac{1}{N_k} \sum_{k_i=k}^N Knn(i), \quad (2.24)$$

onde N_k é o número de vértices com grau k . Desse modo obtemos a distribuição acumulada $Knn(k)$. Depois disso, avalia-se a inclinação da reta do gráfico $k \times Knn(k)$, de modo a determinar o nível de assortatividade da rede.

Outra abordagem é calcular o nível de assortatividade a partir do coeficiente de Pearson, isto é

$$\rho = \frac{(1/M) \sum_{j>i} k_i k_j A_{ij} - \left[(1/M) \sum_{j>i} \frac{(k_i+k_j)A_{ij}}{2} \right]^2}{(1/M) \sum_{j>i} \frac{(k_i^2+k_j^2)A_{ij}}{2} - \left[(1/M) \sum_{j>i} \frac{(k_i+k_j)A_{ij}}{2} \right]^2}, \quad (2.25)$$

onde M é o número total de arestas da rede. Se $\rho > 0$, então a rede é assortativa (AN). Se $\rho < 0$ a rede é dissassortativa (DN). Se $\rho = 0$, então não existe correlação entre o grau dos vértices e a rede é não correlacionada (NA).

O cálculo de Pearson é utilizado para analisar o nível de correlação entre duas variáveis, que neste caso podem ser medidas de centralidade, de acordo com a Equação 2.26. A correlação de duas medidas x e y para uma rede é quantificada a partir dos x_i e y_i que representam os correspondentes valores das medidas para todos os vértices da rede e tanto $\langle x \rangle$ como $\langle y \rangle$ são suas correspondentes médias.

$$\rho(x,y) = \frac{\sum_i (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sqrt{\sum_i (x_i - \langle x \rangle)^2} \sqrt{\sum_i (y_i - \langle y \rangle)^2}} \quad (2.26)$$

Os valores da correlação da Equação 2.26 variam de -1 (correlação negativa) até 1 (correlação positiva), com o valor de 0 indicando que as duas variáveis não são correlacionadas. Uma correlação completa entre duas medidas ($\rho(x,y) = 1$) pode ser considerada como uma situação de redundância de informação, dado que nenhuma informação adicional está sendo agregada. Por outro lado, no caso de medidas com um valor alto de correlação mas não completa, podem agregar informações adicionais na caracterização e separação dos vértices e da rede (COSTA *et al.*, 2007).

2.1.4.2 Medidas de distância

Uma medida que quantifica a eficiência da rede no envio de informações através dos vértices é a proposta por Latora e Marchiori (2001) chamada de eficiência global, a qual é definida

$$E = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{\ell_{ij}}, \quad (2.27)$$

onde a soma leva em conta todos os pares de vértices da rede. Esse método assume que a eficiência para enviar informação entre os vértices i e j é proporcional ao recíproco de suas distâncias. A reciprocidade da eficiência global é a média harmônica das distâncias geodésicas. A média dos caminhos mínimos $\langle \ell \rangle$ é expressada como

$$\langle \ell \rangle = \frac{1}{N(N-1)} \sum_{i \neq j} \ell_{ij}. \quad (2.28)$$

Assim ao calcular-se $h = 1/E$ estaremos evitando o problema de divergência para redes não conectadas que apresenta a Equação 2.28.

2.1.4.3 Modularidade

A modularidade é uma medida que busca encontrar a divisão ótima do número de comunidades na rede. A medida é calculada por uma matriz simétrica quadrada $\mathfrak{E}$ de c comunidades, na qual os elementos ao longo da diagonal principal, $\mathfrak{E}_{i,i}$, representam o número de conexões entre vértices da mesma comunidade e os elementos $\mathfrak{E}_{i,j}$, representam as conexões entre vértices de diferentes comunidades, i e j . Na Equação 2.29 é apresentada a medida de modularidade Q .

$$Q = \sum_i \left[\mathfrak{E}_{i,i} - \left(\sum_j \mathfrak{E}_{i,j} \right)^2 \right]. \quad (2.29)$$

A equação mede a fração de arestas na rede que conectam vértices de uma mesma comunidade, menos o valor esperado da mesma divisão, mas considerando-se conexões aleatórias entre os vértices. Se uma divisão tem menos conexões entre comunidades do que o esperado, $Q = 0$. Se a rede possui comunidades isoladas, $Q = 1$.

Essa medida é utilizada por diversas técnicas para a identificação de comunidades, especialmente em algoritmos divisivos e aglomerativos. Especificamente, Newman (GIRVAN; NEWMAN, 2002; NEWMAN, 2004; NEWMAN, 2006) tem participado ativamente na proposta de algumas técnicas de identificação de comunidades, as quais serão apresentadas a seguir:

Baseado na intermediação das arestas

Girvan e Newman (2002) propõem uma técnica divisiva de detecção de comunidades. As técnicas divisivas se fundamentam na ideia de que ao remover as arestas que conectam os vértices entre as comunidades é possível identificar o conjunto de vértices que compõem cada comunidade. Para a técnica, as arestas que conectam as comunidades são identificadas mediante a centralidade *betweenness*, mas não é estritamente necessário que seja essa medida. Ao remover as arestas com maior *betweenness*, as conexões intra comunidades tendem a sair primeiro, e assim, a rede é decomposta em sub-redes separadas. Sempre que uma aresta for removida é necessário recalculá-la medida de *betweenness* para a rede. O algoritmo executa iterativamente até que todas as arestas sejam removidas, resultando em uma árvore hierárquica

ou dendograma, Figura 7, onde o corte em vermelho é estabelecido pela configuração que gerou o melhor resultado de Q .

Para o cálculo do *betweenness*, os autores sugerem um algoritmo mais rápido de busca em profundidade. O algoritmo tem uma ordem de $O(M^2N)$ operações por iteração, para uma dada rede. Embora essa técnica represente uma boa alternativa para detectar comunidades, apresenta algumas desvantagens. A principal é seu elevado custo computacional como já foi colocado. Desse modo, algumas abordagens têm sido propostas. Entre elas, o mesmo Newman propôs um método guloso baseado na otimização da modularidade, que será discutido a seguir.

Baseado na otimização da modularidade

Newman (2004) propõe um algoritmo de otimização guloso (em inglês *fastgreedy*), categorizado como um método aglomerativo, no qual uma rede de N vértices começa sem nenhuma conexão, com N comunidades iniciais. A cada iteração, duas comunidades i e j que tenham conexão na rede real são escolhidas se fornecerem o maior acréscimo de valor na função objetivo, expressada na Equação 2.30. No cálculo da função é feita uma poda no espaço de busca considerando apenas os pares de vértices entre os quais existem arestas e, desse modo, o tempo de execução do algoritmo é reduzido.

$$\Delta Q_{ij} = 2 \left(\mathfrak{E}_{i,j} - \frac{\sum_j \mathfrak{E}_{i,j} \sum_i \mathfrak{E}_{i,j}}{2M} \right). \quad (2.30)$$

O resultado também pode ser representado como um dendograma, como no exemplo da Figura 7. Cortes em diferentes níveis do dendograma darão divisões com maior ou menor número de comunidades. O melhor corte é aquele que atinja o maior valor de Q . O algoritmo em cada passo possui complexidade $O(M + N)$. Como há no máximo $N - 1$ operações de junções necessárias para construir o dendograma completo, sua complexidade total é $O((M + N)N)$, ou $O(N^2)$ para uma rede esparsa. Posteriormente, Clauset, Newman e Moore (2004) apresentam um novo algoritmo que opera em tempo $O(N(\log(N))^2)$ para redes esparsas, possibilitando a análise de comunidades em redes maiores. Destaca-se que esse método de otimização é mais rápido que o método anterior de *betweenness*.

Baseado em análise espectral

O método está baseado na análise do espectro da matriz de adjacência (NEWMAN, 2006), cujos valores correspondem aos autovalores da rede. A matriz de adjacência pode ser representada como a matriz laplaciana $L = D - A$, ou laplaciana normalizada, onde D representa a matriz diagonal $D_{ii} = \sum_j A_{ij}$, e $D_{ij} = 0$ para $i \neq j$. Newman propõe um método que reformula a modularidade em relação aos autovetores da matriz, denominada matriz de modularidade $B^{(g)}$, isto é:

$$B_{ij}^{(g)} = A_{ij} - \frac{k_i k_j}{2M} - \delta_{ij} \sum_{u \in N(g)} \left[A_{iu} - \frac{k_i k_u}{2M} \right]. \quad (2.31)$$

Figura 7 – Dendograma das comunidades encontradas pelo algoritmo de *fastgreedy* para a rede clube de karatê. As formas dos vértices representam os dois grupos nos quais o clube se separou.

Fonte: Adaptada de Newman (2004).

A ideia é construir a matriz de modularidade e calcular os autovetores para o maior autovalor positivo. Depois, considera-se uma biseção espectral da matriz, $x = 2$, separando os vértices em duas comunidades de acordo com o sinal de cada vértice no correspondente autovetor. Se todos os elementos no autovetor tem o mesmo sinal, isto significa que a rede não possui uma estrutura de comunidade bem definida. Depois disso, o processo é repetido para cada nova comunidade até que a divisão dê como resultado zero ou uma contribuição negativa ao valor da modularidade. A biseção espectral consiste na diagonalização da matriz de modularidade. Se a rede está separada em c componentes desconexos, a matriz terá c autovetores degradados, todos correspondentes ao valor $\bar{0}$.

2.2 Modelos de Propagação

Neste trabalho foram selecionados as principais modelos que representam dinâmicas de processos de propagação na literatura, que incluem a propagação de epidemias, informações ou rumores. A seguir, apresentamos um estudo atual para cada uma delas, abordando o que tem sido estudado e desenvolvido, as novas propostas e os tópicos de pesquisa atuais.

2.2.1 Propagação de epidemias

A modelagem de epidemias é um tema de pesquisa importante para entender como uma doença se espalha entre os indivíduos de um sistema e como pode ser controlada. Considera-se que a doença se propaga a partir do contato entre indivíduos infectados com aqueles ainda suscetíveis à doença. Formalmente, define-se a transmissão de uma epidemia como o sistema em que circula um agente infeccioso através de um padrão complexo de contato, distinguindo-se claramente os componentes do sistema com o processo de transmissão em si (KOOPMAN, 2005). O contato está relacionado com todos os modos e fatores de interação que incrementam o risco de contágio. Como componentes do sistema temos o agente infeccioso, os hospedeiros e a dinâmica interna dos indivíduos infectados, (Figura 8). O processo de transmissão contempla o mecanismo pelo qual os agentes infecciosos saem dos hospedeiros, os meios ou modos pelos quais os agentes são transmitidos, o processo de contato entre os hospedeiros e os indivíduos suscetíveis para que aconteça a transmissão e a evolução do sistema, que pode contar com mutações, novos meios de transmissão, entre outros. Ademais, existe uma hierarquia intrínseca na transmissão da epidemia relacionada com a interação entre os indivíduos do sistema, assim como a dinâmica interna do patógeno em cada indivíduo (indivíduos imunes ou mais ou menos propensos a transmitir/contrair a doença).

A doença se encontra em um estado endêmico quando existe uma fração finita diferente de zero de indivíduos infectados. Quando a proporção de infectados é mais da metade da população, considera-se um estado de pandemia. Se depois de um tempo não há mais indivíduos infectados, pois os infectados estão recuperados (imunizados) e os restantes são indivíduos suscetíveis que nunca foram contagiados, declaramos que a doença foi curada ou contida. Em muitos casos a doença pode chegar a ser contida no sistema, porém, o desafio está em saber quanto tempo será preciso para que isso aconteça, como fazer para acelerar o processo e quantos indivíduos no final serão contagiados.

A evolução e o mecanismo social de contágio de uma doença pode ser conceitualizado em uma grande variedade de modelos matemáticos de processos de propagação e difusão (HETHCOTE, 2006). A modelagem matemática de processos de propagação de epidemias tem seus primórdios no século XVIII, quando Daniel Bernoulli em 1760 investigou o alto índice de morbidez e mortalidade causado pela varíola. Ele propôs um modelo para calcular o número de pessoas recuperadas ou imunes pela doença (HETHCOTE, 2006).

Figura 8 – Arcabouço da transmissão de uma doença. O sistema inclui: (1) os possíveis hospedeiros, (2) o patógeno ou agente infeccioso, (3) os fatores que estimulam a proliferação, (4) as interações de contato entre os indivíduos, e (5) a dinâmica interna da doença nos indivíduos infectados.

Fonte: Adaptada de [Barrat, Barthélemy e Vespignani \(2008\)](#).

Contudo, a modelagem matemática como é feita atualmente começou no início do século XX, quando Hammer, em 1906, propôs um modelo discreto que analisava a recorrência epidêmica do sarampo ([HETHCOTE, 2006](#)). A partir de então as pesquisas aumentaram cada vez mais, até nossos dias ([NEWMAN, 2002b](#); [READ](#); [KEELING, 2003](#)). Os modelos atuais abrangem desde abordagens comportamentais simples até ferramentas estruturais complexas que incluem hierarquias e a heterogeneidade presente nas comunidades e nas camadas sociais. Dentre as diversas abordagens citamos inicialmente ([BARRAT; BARTHÉLEMY; VESPIGNANI, 2008](#)): (i) os modelos homogêneos, no qual os indivíduos interagem de modo aleatório e homogêneo com os demais indivíduos da rede; (ii) estruturas sociais, no qual os indivíduos são classificados de acordo com as informações demográficas; (iii) redes de contato, na qual se reproduz a rede de relações ou interações entre os indivíduos; (iv) modelos de multi-escala, que considera sub-populações acopladas pela transferência de indivíduos entre elas; e (v) modelos baseados em agentes, que simula a movimentação e interações de indivíduos em uma escala mais detalhada, como a representação de uma comunidade, bairro, entre outros.

Também, existem propostas que são combinações das anteriores, como no caso de múltiplas sub-populações que podem incluir ou não estruturas espaciais acopladas pela transferência de indivíduos, e onde cada sub-população é tratada de modo específico dependendo do caso de estudo ([ANDERSON; MAY, 1992](#); [BOLKER; GRENFELL, 1993](#); [GRENFELL, 1997](#); [KEELING; ROHANI, 2008](#)); modelos de multi-escala com diferentes níveis de granularidade (região, país, comunidade de países, etc.) considerando abordagens como redes de fluxo de pessoas ou bens ([RVACHEV; LONGINI, 1985](#); [KEELING; EAMES, 2005](#); [COLIZZA; PASTOR-SATORRAS; VESPIGNANI, 2007](#); [KEELING; ROHANI, 2008](#)); ou abordagens

baseadas em agentes, simulando o processo de contágio em escala individual (CHOWELL *et al.*, 2003; EUBANK *et al.*, 2004).

Um fator importante a se destacar na escolha de um modelo é sua precisão comparada com os casos do mundo real. O modelo deve buscar ser ótimo de modo que também continue sendo simples e com a menor quantidade de parâmetros configuráveis. Desse modo, as redes de contatos permitem um mapeamento preciso entre as redes sociais com suas conexões e o processo de propagação, oferecendo um nível de precisão próximo da realidade (NEWMAN, 2010; DANO *et al.*, 2011). A rede social possui os possíveis caminhos e a vizinhança de contato por onde a doença pode se espalhar. Desse modo, as doenças se propagam entre indivíduos por meio de: (i) vias aéreas, como no caso da influenza ou tuberculoses transmitidas quando duas pessoas em um mesmo recinto fechado respiram o mesmo ar; (ii) ao toque ou contato, como no caso de alguns parasitas; e, (iii) contato sanguíneo ou transmissão sexual. Assim, os padrões de contato podem ser representados como uma rede. Além disso, os modelos baseados em redes também possuem um número baixo de estados e parâmetros configuráveis.

2.2.1.1 Abordagem geral

Existem várias propostas de modelos que abordam a dinâmica de epidemias, com vários estados, parâmetros e características específicas de contato (BAILEY, 1975; ANDERSON; MAY, 1992; DIEKMANN; HEESTERBEEK, 2000; NEWMAN, 2002b; VESPIGNANI, 2011). No geral, a população é dividida em três grupos: suscetíveis (S), indivíduos que não possuem a doença em questão mas podem adquiri-la se expostos a outro indivíduo infectado; infectados (I), indivíduos que possuem a doença e podem transmiti-la; recuperados (R), indivíduos que estão recuperados da doença e possuem imunidade permanente, desse modo não poderão adquirir de novo a doença e nem transmiti-la. Os modelos tradicionais de propagação de epidemias foram introduzidos nos trabalhos (KERMACK; MCKENDRICK, 1927; ANDERSON; MAY, 1992; KEELING; EAMES, 2005; VESPIGNANI, 2011), ou seja, os modelos suscetível-infectado (SI), suscetível-infectado-recuperado (SIR) e suscetível-infectado-suscetível (SIS). Na Seção 2.2.1.2 será explicado com maiores detalhes esses modelos e suas dinâmicas.

A propagação da epidemia acontece com uma probabilidade β quando um indivíduo infectado tem contato com um indivíduo suscetível, $I + S \xrightarrow{\beta} 2I$. Por outro lado, a recuperação do indivíduo acontece de modo espontâneo depois de algum tempo dt , com uma probabilidade μ a cada instante de tempo. Em todo momento a população é constante, com $|I| + |S| + |R| = N$.

Algumas propriedades relacionadas com as redes e com a dinâmica podem ser generalizadas em modelos matemáticos que permitam estudar situações particulares e melhorar o entendimento dos efeitos da propagação, como por exemplo o limiar epidêmico. Este é um conceito epidemiológico fundamental para o entendimento da propagação e está relacionado com o número básico de reprodução R_0 . O parâmetro R_0 indica o número de casos de infectados depois do primeiro indivíduo infectado da rede. Para que a epidemia tenha chance de

se espalhar precisa-se que $R_0 > 1$ (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008). Nesse sentido, a epidemia será capaz de gerar um número maior de infectados em comparação aos que se recuperam, levando a um incremento de indivíduos infectados por fração de tempo. O limiar epidêmico é um conceito essencial em epidemiologia, dado que oferece um marco de referência para o entendimento da evolução do processo de contágio, além de servir como medida para ser minimizada ao tentar conter ou eliminar a doença (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008).

Contudo, a maioria dos modelos assume que a rede é estática durante a propagação, o que difere do fenômeno real, onde a rede pode ser dinâmica e mudar durante a difusão da doença.

2.2.1.2 Modelagem matemática

Considera-se que a população é constante durante o processo de propagação, com $S(t) + I(t) + R(t) = N$, onde $S(t)$ representa o número de vértices suscetíveis no tempo t , e assim para os outros estados. Seleciona-se um vértice da rede como infectado, escolhido aleatoriamente, para dar início à dinâmica. Pode-se definir a densidade de vértices suscetíveis como $s(t) = S(t)/N$ e de modo equivalente para os demais estados. Assim, as densidades iniciais são definidas por $s(0) = (N - 1)/N$, $i(0) = 1/N$ e $r(0) = 0$, onde $\forall t : s(t) + i(t) + r(t) = 1$.

O primeiro modelo de epidemia a ser tratado será o SI, no qual existem apenas dois estados discretos em que podem estar os vértices, com $s(t) + i(t) = 1$. Para os vértices suscetíveis, as chances para que um dado vértice i adquira uma doença é proporcional a média de contatos que ele tem, representado por k_i . Porém, pela hipótese de homogeneidade, tem-se que $k_i \sim \langle k \rangle$. Desse modo, considerando a taxa β de contágio, a probabilidade de que um vértice suscetível adquira a doença em um lapso de tempo Δt é dada por $\beta \langle k \rangle$, com $\Delta t = t_2 - t_1$ e $t_2 > t_1$. Assim, a formulação discreta que define a densidade de vértices infectados no tempo é,

$$i(t + \Delta t) = i(t) + s(t)\beta \langle k \rangle \Delta t, \quad (2.32)$$

onde $s(t)\beta \langle k \rangle$ é a média global da fração de novas infecções no intervalo Δt . Portanto, calculando o limite quando $\Delta t \rightarrow 0$, para descrever assim a taxa de infectados no tempo do modelo e considerando que $s(t) = 1 - i(t)$ tem-se

$$\frac{di(t)}{dt} = \beta \langle k \rangle i(t) [1 - i(t)], \quad (2.33)$$

na qual, o crescimento diferencial no tempo é proporcional com a probabilidade de contágio $\beta \langle k \rangle$ e a densidade de vértices infectados junto com a densidade de vértices suscetíveis, que ainda restam na rede. A solução exata da Equação 2.33 é dada por (BARRAT; BARTHÉLEMY;

VESPIGNANI, 2008):

$$i(t) = \frac{i_0 e^{t/\tau}}{1 + i_0(e^{t/\tau} - 1)}; \quad \tau = \frac{1}{\beta \langle k \rangle}, \quad (2.34)$$

onde i_0 representa a densidade inicial de indivíduos infectados e τ o fator de escala no tempo relacionado com a velocidade de crescimento da curva de infectados.

As formulações apresentadas consideram que a rede é homogênea com $k_i \sim \langle k \rangle$. Para generalizar o modelo para redes heterogêneas de qualquer distribuição de grau, considera-se uma aproximação utilizando blocos ou classes de vértices que possuem o mesmo grau, assumindo-se que os vértices do mesmo grau são estatisticamente equivalentes. Assim, são analisadas classes de vértices com o mesmo grau k , quantificados pela densidade de vértices desse conjunto em cada um dos estados da dinâmica, dada por

$$i_k = \frac{I_k}{N_k}; \quad s_k = \frac{S_k}{N_k}, \quad (2.35)$$

onde N_k representa o número de vértices com grau k da rede, I_k e S_k são o número de vértices com grau k que estão infectados e susceptíveis respectivamente. Desse modo, as médias globais para cada estado são definidas pela distribuição de grau da rede ($P(k)$), isto é

$$i = \sum_k P(k) i_k; \quad s = \sum_k P(k) s_k. \quad (2.36)$$

Considerando a probabilidade de encontrar uma aresta que conecte um vértice de grau k' dado que no outro extremo há um vértice com grau k , tem-se que essa probabilidade $P(k'|k)$ definida por (COSTA *et al.*, 2007; BARRAT; BARTHÉLEMY; VESPIGNANI, 2008) é $P(k'|k) = \frac{k'P(k')}{\langle k \rangle}$. Isto indica que a probabilidade condicional de conexão é proporcional a k' sem depender do grau da condição. Assim, define-se o grau médio da rede como $\langle k \rangle = \sum_{k'} P(k') k'$. Considerando a densidade de vértices infectados de grau i_k , a probabilidade condicional de conexão e o grau médio da rede, tem-se que a probabilidade de um vértice de grau k seja infectado é definida como (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008)

$$\Theta_k(t) = \Theta(t) = \frac{\sum_{k'} (k' - 1) P(k') i_{k'}(t)}{\langle k \rangle}. \quad (2.37)$$

Considera-se $(k' - 1)$ possíveis vizinhos pois o vizinho que transmitiu a doença para o vértice selecionado não é considerado como suscetível. Tem-se que $\Theta_k(t) = \Theta(t)$ dado que a densidade de vizinhos infectados também não depende de k .

Avaliando o conjunto ou classe de vértices de grau k , com probabilidade de propagação proporcional a taxa β vezes o número de vizinhos da classe, a probabilidade de que algum vértice

do grupo não esteja infectado $[1 - i_k]$ e a densidade de vizinhos infectados para dito grupo, a partir da Equação 2.33 tem-se que

$$\frac{di_k(t)}{dt} = \beta k [1 - i_k(t)] k \Theta(t). \quad (2.38)$$

Combinando-se as Equações 2.37 e 2.38 e negligenciando termos de ordem superior $O(i^2)$, obtém-se as equações de evolução para o modelo SI (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008)

$$\begin{cases} \frac{di_k(t)}{dt} = \beta k \Theta(t), \\ \frac{d\Theta(t)}{dt} = \beta \left(\frac{\langle k^2 \rangle}{\langle k \rangle} - 1 \right) \Theta(t). \end{cases} \quad (2.39)$$

Esse sistema de equações diferenciais pode ser resolvido tomando como condição inicial uniforme que $i_k(0) = i_0$ (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008). Calculando-se a prevalência de vértices com grau k e fazendo a média das prevalências considerando $i(t) = \sum_k P(k) i_k(t)$, obtém-se a solução

$$i(t) = i_0 \left[1 + \frac{\langle k \rangle^2 - \langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} (e^{t/\tau} - 1) \right], \quad (2.40)$$

com valor de taxa de escala no tempo para a curva de infectados de

$$\tau = \frac{\langle k \rangle}{\beta (\langle k^2 \rangle - \langle k \rangle)}. \quad (2.41)$$

Desse modo, a partir da Equação 2.41, é possível voltar ao mesmo resultado de τ da Equação 2.34 quando avaliada para uma rede com distribuição de grau homogênea.

De modo a simular o modelo SI, define-se o passo de tempo ou iteração temporal do modelo, o qual consiste em avaliar o estado de cada um dos vértices da rede no tempo t e realizar os contatos com os vizinhos de cada vértice. O algoritmo opera do seguinte modo: primeiro, é selecionado de modo aleatório um vértice da rede como infectado inicial no tempo $t = 0$. Depois, para todo vértice infectado i na iteração temporal t são contatados seus vizinhos suscetíveis e tenta-se passar a doença de acordo com a probabilidade β . Para isso, é sorteado um valor aleatório $x|x \in [0, 1]$ para cada vértice suscetível j , vizinho de i . Muda-se o estado de j para infectado, na iteração temporal $t + 1$, se o valor $x \leq \beta$. O processo continua até que não haja mais vértices suscetíveis na rede.

Como pode-se observar na Figura 9, a proporção de infectados apresenta inicialmente um crescimento exponencial até que é alcançado o estado estacionário do modelo, com $i(t) = 1$. Também, observa-se que dependendo do tipo de rede, e ainda com as mesmas condições iniciais,

Figura 9 – Resposta temporal dos infectados para a dinâmica SI em epidemias, nos modelos de redes ER, WS e BA com $N = 1000$, $\langle k \rangle = 8$, $\beta = 0.1$ e $\mu = 0.2$

Fonte: Elaborada pelo autor.

a dinâmica varia podendo crescer mais rapidamente, como no caso de uma rede sem escala (Figura 9d), ou mais lentamente, como no caso de uma rede próxima a ser regular (Figura 9c).

O comportamento do crescimento nas simulações pode ser explicado analisando a taxa de crescimento τ da Equação 2.41. Quando a rede segue uma distribuição de grau $P(k) \sim k^{-\gamma}$ com $2 \leq \gamma \leq 3$, tem-se que a complexidade da rede diverge, com $C \sim \langle k^2 \rangle \rightarrow \infty$ para redes muito grandes ($N \rightarrow \infty$). Desse modo, a escala de tempo τ tende a ser muito pequena aumentando rapidamente a propagação. Diferentemente das redes do tipo ER ou regulares, cujos valores de complexidade são mais baixos e, por isso, ainda com o mesmo tamanho, a curva de infectados cresce mais lentamente.

Nem todas as doenças são incuráveis e nem todas conferem imunidade aos sobreviventes. Exemplos disso são a tuberculose, gripe, ou vírus de computadores. Nesses casos, os indivíduos infectados não chegam a se recuperar sendo imunes ou continuam doentes para sempre, mas voltam a ser suscetíveis. As formulações descritas podem ser estendidas para o modelo suscetível-infectado-suscetível (SIS). A diferença dos modelos está em que os vértices infectados podem voltar a ser suscetíveis de maneira espontânea, segundo uma taxa de recuperação μ . Desse modo, a probabilidade de que um vértice infectado volte a ser suscetível aumenta a medida que passa o tempo, expressada por $\mu \Delta t$. Portanto, adicionando a nova transição espontânea de estado na

definição da Equação 2.33, a formulação para o modelo SIS é dada por

$$\frac{di(t)}{dt} = \beta \langle k \rangle i(t) [1 - i(t)] - \mu i(t), \quad (2.42)$$

onde o termo $\mu i(t)$ é negativo porque o vértice passa de modo espontâneo a ser suscetível, não dependendo do contato ou grau médio da rede. A Equação 2.42 não possui uma solução exata, a qual pode ser calculada mediante métodos numéricos ou simulações computacionais. Contudo, uma aproximação do comportamento inicial da dinâmica para valores pequenos de t pode ser obtida. A partir de uma aproximação linear expandido os termos da Equação 2.42 e negligenciando os termos de segunda ordem $O(i^2)$, tem-se

$$\frac{di(t)}{dt} = \beta \langle k \rangle i(t) - \mu i(t). \quad (2.43)$$

Desse modo, pode ser definida uma solução à densidade de infectados no tempo a partir da Equação 2.43, isto é (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008)

$$i(t) \simeq i_0 e^{t/\tau}; \quad \tau^{-1} = \beta \langle k \rangle - \mu, \quad (2.44)$$

onde i_0 representa a densidade inicial de vértices infectados na rede, e τ é o fator de escala no tempo da capacidade de propagação. Segundo as análises feitas para o fator τ (ANDERSON; MAY, 1992), observa-se que diferente do modelo SI (Equação 2.34), podem aparecer valores negativos quando $\mu > \beta \langle k \rangle$. Para esses casos, a propagação não atingirá uma fração finita de infectados e tende a ser eliminada da rede. Essa consideração faz referência ao conceito do limiar de propagação, que pode ser reescrito como

$$\tau^{-1} = \mu(R_0 - 1) > 0, \quad R_0 = \frac{\beta \langle k \rangle}{\mu}, \quad (2.45)$$

com $R_0 > 1$ para que a propagação aconteça na rede. A partir disso, obtém-se a expressão do limiar epidêmico para os modelos SIS e SIR homogêneos, considerando apenas as taxas de propagação e recuperação da dinâmica, sendo

$$\frac{\beta}{\mu} > \langle k \rangle. \quad (2.46)$$

A generalização para qualquer distribuição de grau, com a inclusão da recuperação estocástica $\mu i_k(t)$ na Equação 2.38, obtém-se a densidade de infectados segundo a classe k para o modelo SIS, dada por (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008)

$$\frac{di_k(t)}{dt} = \beta k [1 - i_k(t)] k \Theta_k(t) - \mu i_k(t). \quad (2.47)$$

No caso do modelo SIS, existe uma diferença na formulação de $\Theta_k(t)$, pois o vértice infectado pode não estar conectado necessariamente com outro infectado, dado que aquele vértice que lhe passou a doença pode não estar mais infectado, por causa da taxa estocástica de recuperação. Desse modo, são considerados em média k' vizinhos e não $(k' - 1)$ como definido na Equação 2.37. Portanto, a definição no modelo SIS da densidade de vizinhos infectados para vértices de grau k é dada por

$$\Theta_k(t) = \frac{\sum_{k'} (k') P(k') i_{k'}(t)}{\langle k \rangle}. \quad (2.48)$$

O valor estacionário do modelo acontece no limite quando $t \rightarrow \infty$, fazendo com que $di_k(t)/dt = 0$ na Equação 2.47. Assim, calculando o valor estacionário i_k e substituindo na Equação 2.48 obtém-se como resultado uma definição de Θ em função dos parâmetros β e μ (PASTOR-SATORRAS; VESPIGNANI, 2001)

$$\Theta = \frac{1}{\langle k \rangle} \sum_k k P(k) \frac{\beta k \Theta}{\mu + \beta k \Theta}. \quad (2.49)$$

Fazendo a derivada $d/d\Theta$ na Equação 2.49 e avaliando no intervalo $[0, \infty)$ tem-se a expressão do limiar epidêmico do modelo SIS, definido como

$$\frac{\beta}{\mu} \geq \frac{\langle k \rangle}{\langle k^2 \rangle} \quad (2.50)$$

Na Figura 10 são apresentadas algumas simulações do modelo SIS em epidemias para redes do tipo *ER*, *WS* e *BA*, todas com mesmo tamanho e grau médio. As simulações iniciam selecionando um vértice da rede de modo aleatório para ser infectado. Para cada iteração de tempo t , selecionam-se os vértices infectados da rede. Para cada vértice infectado i da rede são contactados seus vizinhos suscetíveis j e tenta-se passar a doença de acordo com a taxa β de propagação, mudando seu estado para o tempo $(t + 1)$. Depois que tentou-se passar a doença para os vizinhos suscetíveis de i , decide-se aleatoriamente se i volta a ser suscetível na seguinte iteração $(t + 1)$, de acordo com a taxa de recuperação μ .

O processo não apresenta uma condição de parada como no caso do modelo SI, pois a dinâmica pode continuar sem nunca terminar os infectados da rede. Contudo, espera-se que seja atingido o estado de equilíbrio estacionário da rede. Desse modo, uma condição pode ser que a variância da densidade de infectados da rede, a partir de um certo tempo, seja menor a um valor ε definido, isto é $\sigma(i(T)) < \varepsilon$, dado um conjunto de iterações temporais contínuas $T = \{t_s, t_s + 1, \dots, t_s + z\}$.

Observa-se que o tipo da rede influencia na dinâmica de propagação do modelo SIS. O estado estacionário é atingido mais rapidamente pela rede do tipo *BA*, do mesmo modo que na dinâmica SI. Isto acontece em concordância com a presença de *hubs* da rede, os quais aumentam

Figura 10 – Resposta temporal para a dinâmica SIS em epidemias, nos modelos de redes ER, WS e BA com $N = 1000$, $\langle k \rangle = 8$, $\beta = 0.1$ e $\mu = 0.2$. A curva vermelha representa a fração de indivíduos infectados e a azul os suscetíveis.

Fonte: Elaborada pelo autor.

o valor do segundo momento da rede, fazendo com que o limiar de propagação (Equação 2.50) diminua. No caso de redes mais homogêneas como a WS da Figura 10c, precisa-se de muitas iterações temporais para que seja atingido o equilíbrio estacionário da rede.

O modelo SIS não assume que a recuperação do vértice pode levar a uma imunidade da doença ou a morte do indivíduo. Por tal razão, o modelo SIR é proposto como alternativa (ANDERSON; MAY, 1992), assumindo que os vértices recuperados, a partir da mesma taxa espontânea μ , não participam mais da dinâmica. Isto é, como se fossem removidos da rede. Assim, o novo estado $R(t)$ é adicionado no modelo. O modelo é regido pelo seguinte sistema de equações diferenciais (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008)

$$\begin{cases} \frac{ds(t)}{dt} = -\beta \langle k \rangle i(t) [1 - r(t) - i(t)] , \\ \frac{di(t)}{dt} = -\mu i(t) + \beta \langle k \rangle i(t) [1 - r(t) - i(t)] , \\ \frac{dr(t)}{dt} = \mu i(t) , \end{cases} \quad (2.51)$$

onde $r(t)$ representa a densidade de vértices recuperados da rede. A razão $dr(t)/dt$ indica a taxa com que os indivíduos infectados da rede se recuperam. Isto também indica que $R(t)$

é um estado absorvente da dinâmica, onde para um tempo $t \rightarrow \infty$, todos os vértices da rede que foram infectados serão recuperados. Ao expandir e linearizar $di(t)/dt$ da Equação 2.51 e negligenciando termos $O(i^2)$, se obtém o mesmo resultado da Equação 2.44 do modelo SIS. Isto acontece pois na fase inicial de propagação, ambos os modelos se comportam e crescem de forma semelhante.

No caso da densidade $r(t)$, considera-se da mesma ordem de grandeza de $i(t)$ na fase inicial da propagação. Assim, a aproximação de $i(t)$ empregada no modelo SIS (Equação 2.44) também vale para o modelo SIR. Ademais, e por causa da similaridade na produção de vértices infectados, o limiar epidêmico do modelo, considerando a homogeneidade na distribuição de grau da rede, é também definido pela Equação 2.46.

Assumindo-se uma distribuição de grau heterogênea, a normalização das densidades por classes de vértices são $s_k(t) = 1 - i_k(t) - r_k(t)$, onde $r_k(t)$ representa a densidade de vértices recuperados com grau k no tempo t . No modelo, a epidemia em algum momento vai terminar com uma densidade de vértices infectados igual a zero. Isto ocorre devido a todos os vértices infectados se tornarem recuperados. Assim, ressalta-se a importância de saber a fração final de recuperados (ou suscetíveis) no final da propagação. Para isso, define-se a expressão $r_\infty = \lim_{t \rightarrow \infty} r(t)$, onde $r(t) = \sum_k P(k)r_k(t)$. As equações do modelo SIR considerando a abordagem compartimental são definidas a seguir (MORENO; PASTOR-SATORRAS; VESPIGNANI, 2002; NEWMAN, 2002b; ná; PASTOR-SATORRAS; VESPIGNANI, 2003).

$$\begin{cases} \frac{ds_k(t)}{dt} = -\beta k s_k(t) \Theta_k(t), \\ \frac{di_k(t)}{dt} = -\mu i_k(t) + \beta k s_k(t) \Theta_k(t), \\ \frac{dr(t)}{dt} = \mu i_k(t), \end{cases} \quad (2.52)$$

onde $\Theta_k(t)$ é a média da densidade de vizinhos infectados para vértices com grau k , a qual é definida segundo a Equação 2.37. Considerando uma distribuição homogênea inicial de infectados $i_k(0) = i_0$, tem-se a formulação para a densidade de suscetíveis $s_k(t) = e^{-\beta k \phi(t)}$, e a função auxiliar $\phi(t)$ é (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008)

$$\phi(t) = \frac{1}{\langle k \rangle} \sum_k (k-1) P(k) r_k(t), \quad (2.53)$$

com a qual, considerando $\phi_\infty = \lim_{t \rightarrow \infty} \phi(t)$, define-se r_∞ , como (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008)

$$r_\infty = \sum_k P(k) (1 - e^{-\beta k \phi_\infty}). \quad (2.54)$$

Para calcular o limiar de propagação é utilizada a função $\Theta_k(t)$ (Equação 2.37), fazendo a derivada $d\Theta_k(t)/dt$, colocando em evidência a Θ e isolando τ , tem-se como resultado (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008)

$$\tau = \frac{\langle k \rangle}{\langle k^2 \rangle \beta - (\beta + \mu) \langle k \rangle}, \quad (2.55)$$

com $\tau > 0$ para que a propagação possa atingir uma fração finita de vértices na rede. Assim, o limiar de propagação do modelo SIR, dependendo das taxas β de contágio e μ de recuperação, é definido por

$$\frac{\beta}{\mu} > \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}. \quad (2.56)$$

Para simular o modelo SIR de epidemia é utilizada a abordagem de processo reativo (RP). Nesta abordagem, são selecionados os vértices infectados da rede durante o tempo t . Para cada vértice infectado, primeiro se tenta passar a doença para todos seus vizinhos suscetíveis, segundo a taxa de contágio β . Depois de contaminar seus vizinhos, o correspondente vértice infectado é avaliado para torná-lo recuperado na seguinte iteração, de acordo com a taxa μ . O processo continua até que não existam mais vértices infectados na rede.

Na Figura 11 são apresentadas algumas simulações do modelo SIR para diferentes tipos de redes artificiais. Observa-se que dependendo do tipo de rede, o processo de propagação muda. Quando a epidemia se espalha em redes do tipo *BA*, é atingido um maior pico de infectados, o qual acontece muito mais rápido em comparação com as outras curvas. Porém, ainda que apresente a maior velocidade inicial na propagação, a fração final de indivíduos infectados foi semelhante a da simulação na rede *ER*. Desse modo, não necessariamente infectar mais pessoas no menor tempo leva a uma maior densidade final de infectados na rede. Por outro lado, no caso da Figura 11c, a configuração dos parâmetros β e μ não foi maior ao limiar epidêmico da rede e, portanto, a propagação não teve sucesso.

2.2.1.3 Influência da estrutura da rede

A topologia, isto é, o padrão de contatos entre os indivíduos, determina as rotas de propagação da epidemia. Alguns trabalhos têm analisado como a estrutura da rede influencia na difusão dos agentes infecciosos (PASTOR-SATORRAS; VESPIGNANI, 2001; KEELING; EAMES, 2005; HETHCOTE, 2006; MILLER; KISS, 2014), buscando entender a dinâmica e o comportamento da propagação para diferentes sistemas complexos. Resultados aplicados em redes de natureza distinta mostram que a propagação da epidemia está relacionada com tipo de rede (KEELING; EAMES, 2005; CASTELLANO; FORTUNATO; LORETO, 2009). Assim, redes que compartilham características semelhantes, como a distribuição de grau, a proporção de triângulos, assortatividade, entre outros, apresentam evolução da epidemia de

Figura 11 – Resposta temporal para a dinâmica SIR em epidemias, nos modelos de redes ER, WS e BA com $N = 1000$, $\langle k \rangle = 8$, $\beta = 0.1$, $\mu = 0.2$, e as respectivas curvas da fração de infectados (vermelho), suscetíveis (azul) e recuperados (verde)

Fonte: Elaborada pelo autor.

forma equivalente. Ademais, redes do tipo sem escala podem não apresentar um limiar epidêmico (PASTOR-SATORRAS; VESPIGNANI, 2001; MELONI *et al.*, 2012) quando são redes muito grandes (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008; VESPIGNANI, 2011). Além disso, os resultados podem variar dependendo da proporção de triângulos na rede (CHEN *et al.*, 2013). No caso da correlação de grau dos vértices, existem evidências que relacionam o potencial de propagação da rede com o quanto os vértices são similares com aqueles que se conectam (NEWMAN, 2002a; ná; PASTOR-SATORRAS; VESPIGNANI, 2003; NEKOVEE *et al.*, 2007). Nesse sentido, a assortatividade exerce um papel importante na difusão da epidemia. Também, a dinâmica possui um papel relevante na propagação. Por exemplo, no modelo SIS, como vírus de computador, foi observado que o limiar epidêmico é nulo (PASTOR-SATORRAS; VESPIGNANI, 2001; KITSACK *et al.*, 2010) para as redes sem escala, infectando uma fração finita de indivíduos, independente da taxa de transmissão.

Outra questão relevante no estudo de epidemias está relacionado com a identificação dos principais espalhadores. A doença pode ser contida vacinando-se de modo seletivo esses indivíduos, ou isolando aqueles que representam um maior risco para que a epidemia se propague. Identificar os melhores propagadores apresenta dois problemas de pesquisa (PEI; MAKSE, 2013): *i*) caracterizar de modo geral os indivíduos que são mais influentes no processo de propagação (KITSACK *et al.*, 2010) e *ii*) encontrar o conjunto de vértices que maximizam o

potencial de propagação da rede (RICHARDSON; DOMINGOS, 2002).

No primeiro caso, para encontrar os vértices mais influentes é feita uma análise da rede em uma escala microscópica. São consideradas as medidas topológicas e dinâmicas dos vértices, classificando-se e selecionando-se os vértices espalhadores de acordo com alguma dessas medidas, para assim ponderar sua capacidade de propagação. Para o segundo problema de pesquisa, investiga-se qual conjunto de vértices iniciais trazem uma maximização do efeito de propagação da epidemia. A primeira vista, pode-se pensar que o problema de encontrar os múltiplos propagadores iniciais é resolvido ao selecionar os vértices que são mais influentes do primeiro caso. Contudo, essa abordagem não resolve o problema, dado que para esses vértices existe uma sobreposição da influência na população (KITSAK *et al.*, 2010) e de fato trata-se de um problema de otimização NP-difícil (KEMPE; KLEINBERG; TARDOS, 2003). No entanto, pode-se adotar soluções heurísticas para determinar os principais propagadores, que podem ser obtidos através de simulações estocásticas.

As redes reais de interação social apresentam uma distribuição sem escala (BARABÁSI, 2001; NEWMAN, 2003; BOCCALETTI *et al.*, 2006; BARABÁSI, 2007), isto é, que a maioria dos indivíduos possuem um número pequeno ou moderado de conexões e os mais destacados da rede são aqueles que tem maior grau (*hubs*). A partir disso, foi encontrado que na propagação de epidemias os *hubs* eram mais propensos a ser infectados e, ademais, representavam os vértices com a maior capacidade de difusão da rede ou mais influentes (PASTOR-SATORRAS; VESPIGNANI, 2001; ALBERT; BARABÁSI, 2002; KITSAK *et al.*, 2010). Outros pesquisadores argumentam que os melhores propagadores são aqueles que apresentam maiores valores de *betweenness* (KITSAK *et al.*, 2010; HÉBERT-DUFRESNE *et al.*, 2013). Esses indivíduos representam aqueles com uma posição estratégica na rede social, intermediando a comunicação entre grupos de vértices, aumentando assim sua influência. Não necessariamente os vértices com maior *betweenness* possuem grau alto. Pode-se tratar de vértices que servem como pontes conectando várias comunidades. Portanto, se esses vértices são vacinados, a difusão da epidemia perderia seu efeito, pois atuam como elos de ligação entre diferentes regiões da rede. Hébert-Dufresne *et al.* (2013) fizeram uma análise local de várias medidas e encontraram que nas comunidades a medida de *betweenness* é um melhor descritor dos propagadores mais influentes.

Ao contrário das pesquisas anteriores, também existem muitas circunstâncias em que os melhores propagadores não são aqueles que apresentam o maior número de conexões (COHEN *et al.*, 2000; PASTOR-SATORRAS; VESPIGNANI, 2001) ou são os mais centrais em termos da medida *betweenness* (FREEMAN, 1979; PASTOR-SATORRAS; VESPIGNANI, 2001; HÉBERT-DUFRESNE *et al.*, 2013). Por exemplo, algumas pesquisas encontraram que os indivíduos mais influentes são aqueles que se encontram no núcleo da rede, identificados ao usar a medida de K-core ou derivações (KITSAK *et al.*, 2010; PEI; MAKSE, 2013; BAE; KIM, 2014). Os pesquisadores argumentam que a posição do vértice propagador, relativa à organização da rede, oferece uma maior influência no processo que simplesmente considerar as características locais

do vértice, como o grau, *betweenness*, entre outros. Por exemplo, ao tratar-se de um *hub* que está na periferia da rede, esse infectará menos indivíduos que um vértice que esteja no núcleo (centro) da rede (KITSACK *et al.*, 2010).

Para o caso de múltiplos infectados iniciais, os vértices com maior grau são mais eficientes que os vértices que estão no núcleo da rede. Isto acontece porque os vértices no núcleo da rede tendem a conectar-se entre si e desse modo, seus efeitos de propagação se anulam. No caso dos *hubs*, no geral nem todos estão conectados entre si e, portanto, conseguem atingir uma maior quantidade de vértices (KITSACK *et al.*, 2010). Medidas estruturais de centralidade podem ser boas para identificar vértices influentes da rede. Porém, não conseguem quantificar o quanto um vértice segundo uma medida central é mais influente que outro da mesma medida (LAWYER, 2012), e nem consideram características próprias da dinâmica (SIKIC *et al.*, 2013).

Novas propostas consideram a combinação de características topológicas junto com a dinâmica de propagação. Por exemplo, a esperança do grau de *cluster* (LAWYER, 2012) quantifica a influência dos vértices considerando a proporção do grau do vértice em relação às comunidades da rede. O grau do vértice em relação a uma comunidade é definido como o número de arestas que se conectam com vértices dentro e fora da comunidade. A medida calcula para um vértice *i*, quantas possíveis comunidades foram infectadas depois de um número *X* de infecções. Os autores aplicaram o método com dados recentes do brote de Ebola em Uganda, e concluíram a partir das características da epidemia, que a doença não atingirá uma propagação a nível global.

Como pode ser observado, não existe um consenso sobre o propagador influente e quais características o definem. Os estudos fazem adaptações ou melhorias nas medidas existentes (CHEN *et al.*, 2013; BAE; KIM, 2014), ou analisam um conjunto de medidas para encontrar a melhor que os descreva (KITSACK *et al.*, 2010; HÉBERT-DUFRESNE *et al.*, 2013; PEI; MAKSE, 2013), propõem novas medidas topológicas (LAWYER, 2012; CHEN *et al.*, 2013), ou analisam tanto a topologia quanto a dinâmica em conjunto (MELONI *et al.*, 2012; SIKIC *et al.*, 2013). Assim, o tema ainda está consolidando-se e há muitas questões a serem provadas e resolvidas.

2.2.2 Propagação de rumores

Os rumores são comuns na sociedade e podem ser tanto benéficos como destrutivos. O mercado de ações, a publicidade e marketing, a adoção de novas tecnologias, a política, entre outras, podem ser consideravelmente afetados pelos rumores. Atualmente, rumores se propagam em redes sociais, tais como Facebook e Twitter, e adquirem maior ou menor força, podendo atingir milhões de usuários. Desse modo, o estudo de modelos de propagação de rumores é interessante para aqueles que desejam entender e se beneficiar desse fenômeno social.

Os rumores podem ser abordados como a propagação de uma ideia ou informação que contamina, ou convence, a mente das pessoas. Por tal similaridade, a propagação de rumores

é muitas vezes associada com o modelo de propagação de epidemias (DALEY; KENDALL, 1964; BETTENCOURT *et al.*, 2006; BARRAT; BARTHÉLEMY; VESPIGNANI, 2008; BORGE-HOLTHOEFER; MORENO, 2012), dado que ambos são processos que se transmitem entre os indivíduos de uma rede. Os rumores, assim como as epidemias, são um exemplo de dinâmica de propagação social no qual o processo de contágio ou difusão considerado é de origem psicológica. O fenômeno de contágio psicológico faz referência aos diferentes aspectos pelos quais um indivíduo está predisposto a adotar e difundir alguma informação, ideia, ou conhecimento. No caso da dinâmica de rumor, os suscetíveis (S) são aqueles indivíduos que ainda não o conhecem ou estão desinformados; os infectados são aqueles que difundem o rumor ou informação, e são chamados de *informantes* ou espalhadores, (I); e os recuperados, são aqueles que conhecem o rumor ou estão informados mas perderam o interesse de difundi-lo, chamados de *contidos*, (R). Além dos estados mencionados, podem ser considerados outros estados como céticos, latentes, entre outros (BETTENCOURT *et al.*, 2006). Porém, não são muito analisados na literatura e por isso o foco de nosso estudo será nos três estados principais mencionados.

Existem algumas diferenças entre os modelos de rumores e epidemias, no nível qualitativo, que são importantes na interpretação, mas não mudam de fato as equações do modelo, e quantitativo, que mudam de fato o comportamento do modelo. Inicialmente, a difusão de uma informação é intencional, pois adquirir ou passar a nova ideia pode levar a uma situação vantajosa. Nesse caso em, relação ao contágio de uma doença, trata-se de um processo passivo por parte dos suscetíveis, pois o indivíduo saudável não está interessado em ficar doente e em hipótese não traz vantagem para ele. Além disso, o processo de adquirir a nova ideia ou informação até o ponto de estar convencido implica em um maior tempo de exposição por uma ou várias fontes de informantes. Do ponto de vista qualitativo, existe uma diferença estocástica no modelo em relação à transição espontânea de estado I para R (DALEY; KENDALL, 1964; MAKI; THOMPSON, 1973). Em rumores, um informante pode perder o interesse de difundir a informação se encontra a outros que já a conhecem. O processo de recuperação deixa de ser espontâneo, como no caso de epidemias e se torna um processo de contato entre os indivíduos. Desse modo, o contato de um informante com um contido ou outro informante faz com que ele possa se tornar contido.

A seguir é apresentada uma revisão da literatura sobre modelos de rumores em redes complexas.

2.2.2.1 Abordagem geral

Existem distintas abordagens utilizadas para a dinâmica de propagação de rumores, que podem gerar diferentes resultados nas simulações. Entre elas está o modelo geral proposto por DK , e suas variantes, como a proposta por MT , as quais tem atraído o maior número das pesquisas na área (NEKOVEE *et al.*, 2007; BARRAT; BARTHÉLEMY; VESPIGNANI, 2008). O modelo DK considera um número inicial $S(0) = N$ de ignorantes, $I(0) = 1$ espalhadores, e

$R(0) = 0$ de contidos.

Considera-se que a rede é homogênea, na qual todos os indivíduos tem a mesma probabilidade de interagir. Na proposta apresentada no modelo *MT*, considerou-se a propagação a partir do contato entre os indivíduos adjacentes, sem estar restrito a alguma distribuição de grau específica para a rede.

A transição de estado, $I + S \xrightarrow{\beta} 2 I$, acontece de acordo com uma probabilidade β no contato de um informante com um vizinho ignorante, a cada iteração. No modelo *MT*, um indivíduo informante torna-se contido com uma probabilidade μ quando entra em contato com um vizinho informante, $I + I \xrightarrow{\mu} R + I$, ou quando tem contato com um vizinho contido, $I + R \xrightarrow{\mu} 2 R$. Na versão original do modelo Daley & Kendall, ao contato de dois informantes, ambos se tornam contidos, com $I + I \xrightarrow{\mu} 2 R$. Em todas as iterações a população é constante com $S(t) + I(t) + R(t) = N + 1$. O número de transições para o estado contido R no modelo é proporcional a $I(t)[I(t) + R(t)]$. De modo a simular a dinâmica *DK*, a propagação acontece em cada iteração sorteando-se uma aresta aleatória da rede. Depois, são avaliados os dois vértices de tal aresta sorteada e define-se os estados segundo as regras de transição.

Analicamente, o modelo *DK* não possui um limiar de propagação, o qual significa que para qualquer valor de β , uma fração finita de pessoas serão informadas (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008), como descrito analiticamente na Equação 2.61 e apresentado a seguir:

$$r_{\infty} = 1 - e^{-(1+\beta/\mu)r_{\infty}},$$

onde μ é a taxa constante de transição para o estado contido R e r_{∞} é o valor estacionário esperado para fração de contidos. Maiores detalhes sobre esse modelo serão discutidos na Seção 2.2.2.4

O modelo proposto por Maki e Thompson (1973) (MT) é uma aproximação ao modelo de Daley e Kendall (1964) (DK), com a diferença que se considera a produção de um contido do conjunto de espalhadores por vez, mas é duplicado o valor da taxa de transição para esse estado (GANI, 2000). Para o caso da dinâmica *MT* existem variações ao tratamento da relação dos informantes com os indivíduos da rede. O processo pode ser de um contato por iteração (*CP*), truncado (*TP*), ou reativo (*RP*) (BORGES-HOLTHOEFER; MORENO, 2012; MELONI *et al.*, 2012). O processo de contato *CP* sorteia para cada informante um vizinho por iteração de tempo e resolve a transição de estado segundo a regra que corresponda. No processo truncado *TP*, em cada iteração, para todos os informantes escolhe-se aleatoriamente um vizinho por vez, definindo-se os estados para o contato segundo corresponda. A seleção continua até que o informante visite todos seus vizinhos ou se torne contido, o que acontecer primeiro. O processo reativo *RP* é menos restrito e costuma ser mais utilizado em propagação de epidemia. Em cada iteração, primeiro os informantes tentam passar o rumor entre seus vizinhos ignorantes, para só

depois, definir se acontece transição de estado com os vizinhos contidos e informantes.

Tanto no modelo de rumor de Daley e Kendall (1964) e a variante de Maki e Thompson (1973), a dinâmica não apresenta um limiar de propagação como no caso de epidemias, pois não se contempla a geração espontânea de contidos nos modelos (NEKOVEE *et al.*, 2007; CASTELLANO; FORTUNATO; LORETO, 2009). Vários trabalhos tem pesquisado a dinâmica de rumor analisando diferentes tipos de redes (MORENO; NEKOVEE; VESPIGNANI, 2004; VESPIGNANI, 2011), com abordagens estáticas ou estocásticas dos modelos (MORENO; NEKOVEE; PACHECO, 2004; NEKOVEE *et al.*, 2007), avaliando o impacto da estrutura na propagação (LI *et al.*, 2014), como por exemplo a existência de vértices mais influentes na rede (KITSACK *et al.*, 2010; BERGE-HOLTHOEFER; MORENO, 2012).

2.2.2.2 Influência da estrutura

A dinâmica da propagação pode ser influenciada pela estrutura da rede e características do modelo de propagação de rumores. Dependendo disso, pode surgir uma região na configuração dos parâmetros que permita a livre propagação, ou contenção do rumor. Inicialmente foi observada a não existência de um limiar de propagação para alguns tipos de redes e dinâmicas (ZANETTE, 2001b; MORENO; NEKOVEE; PACHECO, 2004; NEKOVEE *et al.*, 2007; CASTELLANO; FORTUNATO; LORETO, 2009). Para modelos *MT* e variantes nas redes sem escala, foi encontrado que os *hubs* tendem a se converter rapidamente em espalhadores e pouco depois em contidos (MORENO; NEKOVEE; VESPIGNANI, 2004). Por tal motivo, a fração final de indivíduos informados, ou que conheceram o rumor, no tempo infinito φ_{∞} é menor comparada as redes homogêneas (CASTELLANO; FORTUNATO; LORETO, 2009). Contudo, a propagação nas redes sem escala é mais eficiente no início da dinâmica até atingir o pico máximo de infectados, em comparação com outro tipo de redes (MORENO; NEKOVEE; VESPIGNANI, 2004; CASTELLANO; FORTUNATO; LORETO, 2009).

No caso das redes de pequeno mundo, foram realizadas análises de simulações com o modelo *MT*, considerando-se redes estáticas e dinâmicas (ZANETTE, 2001a; ZANETTE, 2001b). Os estudos mostraram que, ao variar o nível de aleatoriedade no modelo *WS*, acontece uma transição crítica relacionada com o limiar de propagação. Esse modelo interpola de redes regulares até redes com propriedades de pequeno mundo. Dada uma configuração de parâmetros e considerando que a rede é regular, o rumor atinge uma pequena parte da vizinhança do vértice propagador. Conforme se aumenta a probabilidade de reconexão e a rede adquire características de redes de pequeno mundo, foi constatada uma difusão mais rápida e fácil do rumor (WATTS; STROGATZ, 1998a), atingindo uma fração significativa dos indivíduos. Redes com distribuição exponencial ou sem escala também possuem propriedade de mundo pequeno, mas a dinâmica pode apresentar particularidades entre elas (MORENO; NEKOVEE; PACHECO, 2004). Logo, a distribuição do grau exerce uma influência sobre o processo de propagação de rumores.

Considerando as medidas topológicas da rede, autores como Xulvi-Brunet e Sokolov

(2005) argumentam que na prática, o processo de propagação se desenvolve em redes homogêneas e muitos efeitos acontecem nessa situação. Os autores propõem um modelo para alterar o nível de correlação de grau dos vértices usando um parâmetro $-1.0 \leq \alpha \leq 1.0$, resultando em redes assortativas (*AN*, com correlação positiva na distribuição de grau com $\alpha > 0$), não correlacionadas (*NA*, sem correlação na distribuição de grau com $\alpha = 0$) e dissassortativas (*DN*, com correlação negativa na distribuição de grau com $\alpha < 0$), segundo o nível desejado. Para valores maiores de β , a proporção de propagadores e de contidos aumenta nas redes sem escalas quando elas são assortativas (NEKOVEE *et al.*, 2007). Entretanto, se as redes apresentam níveis muito altos de assortatividade, a propagação do rumor termina rapidamente, pois os *hubs* ao estarem conectados entre si se tornam rapidamente contidos, dividindo a rede e interrompendo a propagação.

Podemos nos questionar se nos modelos de rumor (*DK* ou *MT*) também existem vértices que desempenhem um papel importante na propagação, como acontece no caso de epidemias (Seção 2.2.1.3). Pesquisas realizadas em versões estocásticas do modelo *MT*, em redes sem escala, observaram uma interrelação entre a topologia e a dinâmica de rumor (MORENO; NEKOVEE; VESPIGNANI, 2004; MORENO; NEKOVEE; PACHECO, 2004). Os autores encontraram que a escolha do informante inicial nesse modelo não interfere no resultado final de contidos ou informados na rede. De modo mais abrangente e contrário aos resultados obtidos em epidemias (KITSACK *et al.*, 2010), Borge-Holthoefer e Moreno (2012) também demonstram que não há prova local de que existam vértices com características topológicas em comum que sejam mais influentes que outros. Os pesquisadores realizaram simulações nas abordagens SIR e SIS considerando modelos de propagação *DK* e *MT*. Como resultado, observaram que o número final de contidos na rede, a partir dos mesmos parâmetros, é estatisticamente equivalente sem importar o vértice inicial. De modo mais pontual, mostraram que não existe diferença significativa ao considerar como propagadores os vértices com maior grau, ou aqueles que pertencem ao núcleo da rede, como foi encontrado em epidemias (KITSACK *et al.*, 2010; PEI; MAKSE, 2013; BAE; KIM, 2014). Pelo contrário, os vértices com maior grau ou nível de *k*-core funcionam como curto-circuito no processo de propagação, desempenhando mais o papel de *firewall* na rede. Em outras palavras, constatou-se que a dinâmica potencializa que os vértices mais centrais da rede se convertam rapidamente em contidos (MORENO; NEKOVEE; VESPIGNANI, 2004). Nesse sentido, a dinâmica de rumor não favorece a presença de vértices melhores propagadores, pois vértices centrais não são influentes na dinâmica.

Entretanto, considerando o custo do processo de difusão, os resultados sugerem que é mais econômico começar a propagação a partir dos *hubs* da rede, pois a medida que o grau do primeiro espalhador aumenta, o tempo para atingir o valor assintótico de informados diminui. Assim, ao aumentar a velocidade de difusão consegue-se que mais indivíduos saibam o rumor em um menor tempo. Adicionalmente, os vértices mais centrais estão melhor informados e cientes dos rumores que se propagam na rede. Contudo, nesses trabalhos foram assumidas redes sem correlação de grau, o qual não concorda com a característica marcante nas redes sociais, que são no geral assortativas (NEWMAN, 2002a; BARABÁSI, 2009).

2.2.2.3 Modelos estendidos

A propagação de rumores é um tópico de estudo relativamente recente em comparação com as pesquisas em epidemias. Nos últimos anos teve um aumento considerável nos estudos relacionados com rumores, dado o auge das redes sociais e o potencial de difusão de informação que possuem essas redes. Contudo, há campos de pesquisa em epidemias que também poderiam ser abordados em rumores, mas, que por enquanto não foram reportados na literatura. Exemplo disso são: a propagação de vários rumores na mesma rede, analisando situações de concorrência ou cooperação; modelos de rumores aplicados em redes adaptativas, meta-populações ou multicamadas; e inclusive, o fenômeno como a evolução do rumor enquanto se propaga, como no caso das fofocas ou a mudança da informação.

Nesse sentido, nesta Seção colocamos o conjunto de diversos estudos que ampliam os modelos de rumores, Seja na dinâmica do modelo quanto na configuração de novos parâmetros na rede. Os trabalhos analisam processos semelhantes às situações reais que caracterizam os rumores. Abordagens que consideram a capacidade de memória ou esquecimento na rede (DODDS; WATTS, 2004; ZHAO *et al.*, 2011), estratégias de não redundância na propagação (WU; HUBERMAN, 2007; CIMINI *et al.*, 2011), indivíduos que sabem o rumor por fontes independentes (LI *et al.*, 2014) e a adaptação da rede durante processo de difusão (SANTOS; PACHECO; LENAERTS, 2006; PACHECO; TRAULSEN; NOWAK, 2006; LIU; ZHANG, 2014) são exemplos de modelos estendidos.

Li *et al.* (2014) consideram que podem aparecer de modo espontâneo informantes independentes em posições aleatórias da rede. Eles argumentam que os indivíduos podem conhecer o rumor por outros meios ou por outras redes que usam. Os autores propõem um modelo que introduz o parâmetro σ de aparição de informantes independentes na rede por unidade de tempo. A introdução dos informantes independentes reflete o fato de que um indivíduo pode pertencer a várias redes e possui diferentes fontes de informação. Nas análises, consideraram o modelo SI em redes Erdős-Rényi (ER), Barabási-Albert (BA) e sem escala (SF). No estudo foi encontrado que os informantes independentes contribuem na velocidade e no alcance da propagação. Isso acontece pelo fato de tais informantes poderem aparecer em regiões afastadas da rede, onde a propagação convencional levaria mais tempo para chegar até lá, ou talvez, nem conseguiria chegar. Porém, foi verificado que adicionar uma grande quantidade de informantes independentes não traz uma melhora significativa na eficiência da difusão do rumor. Por outro lado, os autores argumentam que, a média do grau ou conectividade da rede impacta e influencia positivamente na propagação do rumor.

No caso de redes adaptativas e não redundantes, Liu e Zhang (2014) propõem um modelo que recombina as conexões dos vértices informantes, selecionando os vizinhos que lhe ofereçam uma maior recompensa segundo uma função de Fermi (SANTOS; PACHECO; LENAERTS, 2006; PACHECO; TRAULSEN; NOWAK, 2006). A função de Fermi forma parte do ferramental da física estatística, e se usa para descrever a probabilidade que um elétron tem de ocupar um

estado de elétron-energia, segundo seu nível de energia. Na abordagem adaptativa proposta, o vértice informante rompe a conexão com um ignorante quando a recompensa é menor à hipótese de conectar-se aleatoriamente com algum dos seus vizinhos ignorantes de segunda ordem. Também, adotaram a propriedade de não redundância para os informantes durante a dinâmica. Essa propriedade significa que um indivíduo não vai tentar transmitir ou convencer seu vizinho de alguma informação por mais de uma vez. Por tal motivo, o informante se torna contido depois de tentar transmitir o rumor para todos seus vizinhos. No estudo foram feitas extensas simulações considerando-se o modelo SIR reativo (*RP*) sobre vários tipos de redes. O rumor propagou de modo mais rápido e abrangente no modelo proposto, superior aos resultados de versões estáticas. Também, encontraram que o início da dinâmica é fundamental para o processo de propagação, pois define o alcance e a velocidade de difusão do rumor.

Outro possível modelo considera que o indivíduo perca o interesse de transmitir uma informação. Entre os vários motivos temos que o rumor deixou de ser interessante, está obsoleto, não parece muito convincente, ou a pessoa simplesmente o esqueceu. O fenômeno de esquecimento do rumor é considerado uma característica importante a ser considerada nos modelos, e tem efeito direto na existência de um limiar de propagação, que pode determinar o fim da disseminação (NEKOVEE *et al.*, 2007; GU; LI; CAI, 2008; ZHAO *et al.*, 2011). Assim, as regras do encontro de um informante com alguém que conhece o rumor não são o único modo de parar a propagação, também o esquecimento desempenha um papel importante, muito mais se a taxa for alta. A regra por contato para a transição de informantes para contidos, nos modelos tradicionais *DK* ou *MT*, remove o limiar de propagação para o modelo SIR (CASTELLANO; FORTUNATO; LORETO, 2009). Ao contrário, quando considerada a taxa de esquecimento na rede, a dinâmica do rumor se assemelha ao comportamento de epidemias, em consideração ao limiar epidêmico e a fração final de infectados (NEKOVEE *et al.*, 2007). Ainda mais, o mecanismo de esquecimento enfraquece a influência da propagação do rumor (ZHAO *et al.*, 2011).

Também é interessante mencionar a proposta de contágio social de Dodds e Watts (2004). A proposta introduz o conceito de memória no modelo SIR, tomando em conta o efeito de ser exposto em repetidas ocasiões pelo rumor. Os autores consideram que para adotar uma nova ideia ou estar convencido de algum rumor é preciso de um tempo e de ser exposto a mais de uma fonte de informação. Entretanto, o mecanismo de memória apresentado não necessariamente muda a formulação matemática do modelo, ou a dinâmica de propagação. A proposta de memória é importante e poderia ser considerada em outras abordagens ou situações reais na difusão de informação.

Outra abordagem considera que os informantes não estão dispostos a difundir o rumor em todo momento, ou também, que os ignorantes podem não ter interesse de espalhá-lo. Ambos os casos foram propostos e estudados separadamente (BORGES-HOLTHOEFER *et al.*, 2012), buscando aproximar alguns fatos reais com os resultados teóricos em redes. No primeiro modelo (BORGES-HOLTHOEFER *et al.*, 2012), os vértices possuem uma probabilidade α_i que afeta

seu comportamento ativo como informante. Isso significa que os informantes só espalham o rumor quando se sintam ativos, dependendo da probabilidade de atividade α_i de cada um. Como consequência, a transição de ignorantes para o estado informante acontece mais devagar. No trabalho foram exploradas as funções de probabilidade uniforme, exponencial e lei de potências, para representar a distribuição de atividade dos vértices. A distribuição heterogênea de atividade nos informantes foi mais relevante para os efeitos da propagação. A distribuição pode ser aleatória e independente da topologia, ou dependente com a distribuição de grau, com $\alpha_i = k_i/k_{max}$. No caso aleatório, para valores altos de α , o rumor tem poucas possibilidades de infectar uma fração significativa da população, pois os informantes se transformam rapidamente em contidos, ao contrário do que acontece com valores baixos de α . Para a distribuição de atividade correlacionado com o grau do vértice, observaram uma forte independência da dinâmica com os valores de α .

No caso da apatia, em que os ignorantes não estão interessados em receber o rumor (BORGE-HOLTHOEFER; RIVERO; MORENO, 2012), considera-se que os indivíduos ignorantes podem passar diretamente ao estado contido, sem participar na dinâmica de difusão do rumor. Esse tipo de comportamento é comum em redes sociais como *Twitter*, onde algumas pessoas recebem mensagens e não a difundem na sua rede. No modelo é adicionada a probabilidade, p , de que um ignorante esteja interessado no assunto e decida difundir-lo. Assim, o ignorante contatado por um informante possui uma probabilidade βp de se tornar informante, e a probabilidade $\beta(1 - p)$ de se tornar diretamente em contido. As análises das simulações sugerem que quando os ignorantes tem uma probabilidade baixa de estarem interessados no rumor, os resultados são mais próximos aos observados nas redes sociais reais.

Os trabalhos apresentados nesta Seção são exemplos que evidenciam como os modelos clássicos de rumor são pouco precisos na aproximação com a realidade. Foi apresentado com alguns exemplos que com pequenas mudanças no modelo ou na rede, baseadas em observações empíricas, se pode aumentar significativamente o nível de realismo do modelo e no seu comportamento.

2.2.2.4 Modelagem matemática

Como foi observado anteriormente, o comportamento da propagação da epidemia depende da rede e da dinâmica do modelo. Para muitas aplicações (BOCCALETTI *et al.*, 2006), o interesse pode não ser prevenir ou deter a disseminação, mas estimulá-la de modo que seja o mais rápida e eficiente possível. Por exemplo, aplicações *Webbots* para disseminação, mapeamento e recuperação de dados na Internet, ou campanhas publicitárias utilizando estratégias virais (*viral marketing*).

As abordagens de propagação de rumores buscam desenvolver modelos “epidemicos” em que, a disseminação de informação a partir de algum vértice da rede atinja o maior número de vértices possível. Qualitativamente, uma das principais diferenças com a modelagem de

epidemias é que no modelo pode-se projetar diferentes regras e condições visando atingir a melhor propagação, ao invés de estar sujeito a modelar processos de contágio de doenças existentes (BOCCALETTI *et al.*, 2006). Além disso, os modelos de epidemia têm sua evolução baseada em taxas estocásticas no tempo, com exceção da fase de infecção, não dependendo do estado de seus vizinhos. Os modelos de rumor baseiam-se no contato direto entre os indivíduos da rede, os quais pertencem a diferentes classes.

Os modelos de rumor consideram geralmente os estados suscetíveis ou ignorantes do rumor (S), informantes (I) e recuperados (R), que são os informantes que pararam de propagar o rumor. O modelo padrão de rumor é o modelo Daley e Kendall (1964) (DK). A principal modificação introduzida no modelo DK consiste em que os informantes passam a ser contidos dependendo dos contatos que fazem e não por uma probabilidade no tempo. Cada um dos N vértices da rede só pode estar em um único estado, isto é para cada vértice i , $\{I_i = 1 \mapsto i \in I\}$, com $S_i + I_i + R_i = 1$. Também, através do tempo a condição do estado continua, com $S(t) + I(t) + R(t) = N$, sendo $S(t)$ o número de vértices no estado S no tempo t , e assim para os outros estados.

Uma variante do modelo DK é o modelo proposto por Maki e Thompson (1973) (MT). A diferença entre os dois modelos está na produção de contidos durante a dinâmica. Consideram-se as taxas β de propagação e μ de contenção para fazer as transições de estado, as quais podem ser definidas entre $[0, 1]$. As regras de mudança de estado do modelo MT, segundo o contato entre os vértices da rede, são sumarizadas a seguir:

$$\left\{ \begin{array}{l} I_i + S_j \xrightarrow{\beta} I_i I_j, \\ I_i + R_j \xrightarrow{\mu} R_i R_j, \\ I_i + I_j \xrightarrow{\mu} R_i I_j, \end{array} \right. \quad (2.57)$$

onde i e j são vértices adjacentes na rede e o operador “+” simboliza o contato entre eles. Observa-se que os informantes desempenham o papel ativo no modelo de rumor. Quando um vértice informante i entra em contato com um vizinho ignorante j , o vizinho se torna um novo informante segundo a taxa β . Quando o informante contacta um vizinho contido, o informante pode tornar-se contido com uma taxa μ , isto é, considera-se que o rumor já é muito conhecido. No caso em que o informante i contacta um vizinho informante j , o vértice i se torna contido de acordo com μ , ao considerar que o rumor não é mais novidade. A diferença do modelo DK está em que ao encontro de dois informantes, os dois podem tornar-se contidos, isto é $I_i + I_j \xrightarrow{\mu} R_i R_j$.

O modelo será inicialmente descrito assumindo-se a hipótese de homogeneidade na distribuição de grau da rede, onde $k \sim \langle k \rangle$. Define-se a função de densidade de vértices ignorantes no tempo como $\psi(t) = S(t)/N$, e de modo equivalente para os outros estados, sendo $\psi(t) +$

$\phi(t) + \varphi(t) = 1$. A formulação no campo médio para a evolução das três densidades do modelo são descritas a seguir (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008)

$$\frac{d\psi(t)}{dt} = -\beta \langle k \rangle \psi(t)\phi(t), \quad (2.58)$$

$$\frac{d\phi(t)}{dt} = \beta \langle k \rangle \psi(t)\phi(t) - \mu \langle k \rangle \phi(t)[\phi(t) + \varphi(t)], \quad (2.59)$$

$$\frac{d\varphi(t)}{dt} = \mu \langle k \rangle \phi(t)[\phi(t) + \varphi(t)]. \quad (2.60)$$

As expressões $\psi(t)\phi(t)$ e $\phi(t)[\phi(t) + \varphi(t)]$ são a força de interação em massa da população segundo as respectivas regras de contágio e contenção do modelo. Na Equação 2.58 temos que a densidade de suscetíveis decai na velocidade com que são convencidos mais informantes. Isto é, o crescimento da densidade de informantes aumenta a uma taxa proporcional de convencimento β , a média de contatos de cada vértice $\langle k \rangle$, e força das interação ($\psi(t)\phi(t)$) entre os suscetíveis e os informantes da rede. Por outro lado, a densidade de informantes cai no tempo convertendo-se em recuperados a uma taxa proporcional a $\mu \langle k \rangle$ vezes a densidade de indivíduos que tem conhecimento do rumor, isto é $1 - \psi(t) = \phi(t) + \varphi(t)$.

O sistema de equações diferenciais pode ser resolvido analiticamente ao se considerar os estados estacionários do modelo. O modelo atinge o estado estacionário quando $t \rightarrow \infty$, desse modo tem-se que $\phi_\infty \equiv \lim_{t \rightarrow \infty} \phi(t) = 0$ e $\varphi_\infty \equiv \lim_{t \rightarrow \infty} \varphi(t)$. A partir de φ_∞ sabe-se se o rumor atingiu uma fração finita da rede ($\varphi_\infty > 0$) e quanto foi essa densidade final de recuperados.

Considerando que o rumor inicia a partir de um único vértice informante, com $\psi(0) = (N-1)/N$, $\phi(0) = 1/N$ e $\varphi(0) = 0$, e que $\varphi_\infty + \psi_\infty = 1$, tem-se que $\varphi_\infty = 1 - \int_0^\infty \psi(t)dt$. Desse modo, calcula-se a função obtendo (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008)

$$\varphi_\infty = 1 - e^{-(1+\beta/\mu)} \varphi_\infty \quad (2.61)$$

Observa-se que a equação é transcendental pois depende dela mesma para ser calculada. Assim, pode-se reformulá-la considerando $\varphi_\infty = F(\varphi_\infty)$, onde $F(x) = 1 - \exp[-(1+\beta/\mu)x]$ é uma função monotônica contínua e crescente, com solução trivial $F(0) = 0$ e $F(1) < 1$. Portanto, uma solução diferente de zero existe quando $dF(x)/dx > 1$, isto é, a primeira derivada de F . Calculando-se a primeira derivada e resolvendo a desigualdade, tem-se que

$$\frac{\beta}{\mu} > 0, \quad (2.62)$$

o qual implica que para quaisquer valores de taxa de propagação β ou de contenção μ o rumor sempre se propagará para uma fração finita da população. Portanto, não é evidenciado para

redes homogêneas uma condição mínima de limiar de propagação, diferentemente do analisado em epidemias. Isto é devido pela diferença na produção de recuperados no modelo, a qual não está relacionada com uma taxa isolada no tempo, como no caso de epidemias, e sim pelo contato entre os indivíduos da rede.

Para a formulação do modelo com distribuição de grau heterogênea, consideram-se a abordagem compartimental dividindo a rede em grupos ou classes de vértices, onde cada grupo possui o mesmo grau k . As equações diferenciais sobre a evolução das densidades do modelo são (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008)

$$\begin{aligned}\frac{d\psi_k}{dt} &= -\beta k \psi_k(t) \sum_{k'} \phi'_k(t) \frac{k'-1}{k'} P(k'|k), \\ \frac{d\phi_k}{dt} &= \beta k \psi_k(t) \sum_{k'} \phi'_k(t) \frac{k'-1}{k'} P(k'|k) - \mu k \phi_k(t) \sum_k [\phi'_k(t) + \varphi'_k(t)] P(k|k), \\ \frac{d\varphi_k}{dt} &= \mu k \phi_k(t) \sum_k [\phi'_k(t) + \varphi'_k(t)] P(k|k).\end{aligned}\tag{2.63}$$

Assim, para redes com topologias mais complexas, com distribuição de grau heterogêneo, é preciso resolver o sistema de modo numérico. Por exemplo, utilizando abordagens como simulações de Monte Carlo.

ABORDAGEM GERAL

Neste capítulo resumizamos os aspectos gerais considerados em cada um dos problemas de pesquisa reportados neste documento. Na Seção 3.1 discutimos sobre os modelos comumente considerados na literatura para as análises de propagação de informação, apresentamos um comparativo qualitativo entre eles e definimos a modelagem adotada na tese. Na Seção 3.2 definimos a configuração geral das simulações e o algoritmo de Monte Carlo empregado. Por último, na Seção 3.3 descrevemos as redes de dados reais utilizadas neste trabalho.

3.1 Modelo de difusão adotado

Os modelos de propagação de informação são considerados equivalentes aos modelos de difusão de epidemias, no sentido de que os indivíduos são “infectados” com uma ideia ou informação (CASTELLANO; FORTUNATO; LORETO, 2009). Nesse sentido, os modelos matemáticos em sociologia assumem que o processo de propagação é um processo de ativação em cascata que ocorre entre vértices ativados e não ativados. Tal cascata consiste na ativação sucessiva de vértices que conhecem a informação e, subsequentemente, ativam outros vértices de acordo com uma regra de ativação. Estes modelos seguem uma função monótona crescente, na qual os vértices que foram ativados não podem ser desativados (KEMPE; KLEINBERG; TARDOS, 2003; KEMPE; KLEINBERG; TARDOS, 2015; GUILLE *et al.*, 2013; PORTER; GLEESON, 2016). Existem duas principais abordagens, o *IC* e o *LT*. Nas propostas de modelos *IC*, a regra de ativação é baseada em uma probabilidade associada a cada aresta da rede, enquanto que nos modelos *LT*, um limiar de influência é adicionalmente associado aos vértices, o qual representa a fração mínima de vizinhos ativos necessária para a respectiva ativação.

Os modelos de *IC* empregam uma abordagem centrada nos propagadores (GUILLE *et al.*, 2013) em que sempre que um vértice se torna ativo, ele tenta ativar seus vizinhos de acordo com as probabilidades associadas a cada aresta e depois para. Por outro lado, os modelos *LT* usam

uma abordagem centrada nos susceptíveis (GUILLE *et al.*, 2013), na qual vértices inativos se autoativam uma vez que uma determinada fração de vizinhos ativos é excedida. Uma modelagem genérica para as duas abordagens foi proposta por Kempe, Kleinberg e Tardos (2003), na qual os autores unificam o *IC* e *LT* em uma mesma modelagem conceitual, sugerindo que os resultados em algum deles (*IC* ou *LT*) podem ser estendidos para o outro. Por esta razão, vamos considerar apenas o modelo *IC* na subsequente análise.

A diferença entre o modelo *SIR* e o *IC* está na transição para o estado contido ou recuperado (propagadores inativos) (GUILLE *et al.*, 2013), a qual acontece instantaneamente na seguinte iteração em *IC*. Em realidade, o *IC* pode ser entendido como uma variação do modelo *SIR*. Em primeiro lugar, os vértices inativos são os correspondentes susceptíveis no modelo *SIR*. Ambos os modelos empregam uma abordagem de contato centrada nos propagadores, o que significa, o contágio ocorre dada a interação dos propagadores. No modelo *IC*, os propagadores tentam somente uma vez passar a informação, o que é equivalente ao modelo *SIR* com parâmetro $\mu = 1$. Ambos os modelos são funções monotônicas crescentes e podem considerar probabilidades heterogêneas no processo de propagação (BUONO *et al.*, 2013; PORTER; GLEESON, 2016; VEGA-OLIVEROS; COSTA; RODRIGUES, 2017b). Contudo, modelos de propagação como os baseados em cascata e derivações fazem bastantes suposições (BONCHI, 2011). Assim um ponto importante a ser analisado é qual modelo de propagação descreve melhor o fenômeno real e quais dessas suposições são mais ou menos realistas.

Consideramos aqui que os modelos de rumores são os mais adequados para descrever a evolução da difusão de informação em redes sociais, porque a transição de estados acontece apenas através do contato entre pares, diferente dos modelos de epidemia ou sociológicos, nos quais a transição de infectados para outros estados é espontânea.

O modelo de rumor proposto por Maki e Thompson (1973) considera que a informação é transmitida através de contatos entre usuários. O indivíduo que conhece o rumor tenta informar a seus vizinhos com uma probabilidade β . Quando o contato é realizado entre dois indivíduos informados, um deles se torna contido de acordo com a probabilidade μ . Nesse caso, o sujeito para a propagação porque considera que a informação perdeu seu valor de “novidade”. Outro ponto importante é que o contágio acontece através do contato entre os indivíduos, o que significa que a estrutura da rede impacta mais na dinâmica. Além disso, o modelo considera algumas situações sociais reais, como múltiplas opiniões sobre a informação, isto é, positivas, hesitantes ou negativas (STIEGLITZ; DANG-XUAN, 2013; HAN *et al.*, 2015), como na modelagem de rumores em que os propagadores se comportam como indivíduos com opiniões positivas e os contidos com negativas. Finalmente, os propagadores tentam repetidamente convencer da informação a seus vizinhos, o qual é uma representação do fenômeno de reforço social (TABACCHI *et al.*, 2017).

Resumimos a seguir alguns pontos importantes da modelagem de Rumore (MAKI; THOMPSON, 1973; BERGE-HOLTHOEFER; MORENO, 2012; HAN *et al.*, 2015; PASTOR-

(SATORRAS *et al.*, 2015) para propagação de informação em redes:

- O contágio acontece através do processo de probabilidade de contato entre os indivíduos, o que significa que a estrutura da rede impacta mais na dinâmica;
- No modelo de rumores, os propagadores param ao decidir que a informação parou de ser interessante, quando considera que o rumor é amplamente conhecido ou sem novidade, de acordo com a interação com seus contatos. Nos modelos de epidemia e em cascata, os indivíduos têm a mesma tendência de parar de disseminar a informação;
- As pessoas podem ter diferentes opiniões sobre uma informação, isto é, positivas, hesitantes ou negativas, como no caso de rumores em que os propagadores se comportam como indivíduos com opinião positiva e os contidos com negativa, além de modelos de rumor mais elaborados terem sido propostos (HAN *et al.*, 2015);
- Os propagadores tentam em repetidas ocasiões propagar as informações para seus vizinhos, o qual está em concordância com fenômeno de reforço social.
- Além disso, os modelos *IC* ou *LT* assumem uma propagação do tipo radiodifusão. Isso significa que os propagadores se comportam como se têm que escolher entre informar imediatamente a todos os seus contatos ou se abster de fazer qualquer coisa, e assim por diante para todos os indivíduos ativados.

Por tais motivos, adotamos o modelo de rumor por seu potencial de descrever melhor o fenômeno de propagação de informações em redes sociais.

3.2 Configuração das simulações

Muitas abordagens têm sido propostos para modelar a dinâmica de propagação de informação e rumores em redes (MORENO; NEKOVEE; PACHECO, 2004; BARRAT; BARTHÉLEMY; VESPIGNANI, 2008; CASTELLANO; FORTUNATO; LORETO, 2009; BORGEHOLTHOEFER *et al.*, 2012). Esses modelos analíticos fazem suposições sobre a estrutura da rede, tais como a correlação ou distribuição de graus, compartimentos ou classe de vértices com as mesmas probabilidades de conexão, mistura homogênea ou campo médio heterogêneo, entre outras. Apesar de todos eles afirmar que suas soluções numéricas concordam com os resultados das simulações de Monte Carlo (MC), optamos por empregar diretamente o método MC.

Em termos computacionais, consideremos uma população de N vértices constantes no tempo. Cada vértice pode estar apenas em um estado, ou seja, $I_i(t) = 1$ se $i \in I$, caso contrário $I_i(t) = 0$, e adicionalmente $S_i(t) + I_i(t) + R_i(i) = 1$. A fração macroscópica de ignorantes ($\psi(t)$), informantes ($\phi(t)$) e contidos ($\varphi(t)$) cumpre sempre que $\psi(t) + \phi(t) + \varphi(t) = 1$, onde a densidade é calculada como $\psi(t) = \frac{1}{N} \sum_{i=1}^N S_i(t)$, e semelhante para as outras densidades. Assumimos que a

Algoritmo 2 – Algoritmo de Monte Carlo para propagação de informação**Entradas:** $G, S(0), I(0), R(0)$,

```

1: Inicializa :  $t \leftarrow 0$ 
2: repita
3:    $t \leftarrow t + 1$ 
4:    $S(t) \leftarrow S(t - 1), I(t) \leftarrow I(t - 1), R(t) \leftarrow R(t - 1)$ 
5:   para todo vértice  $i$  tal que  $i \in I(t - 1)$  faça
6:     para todo vizinho  $j$  de  $i$  faça
7:       // dinâmica de contato segundo o processo adotado
8:       se  $j \in S(t - 1)$  &  $i$  convence  $j$  com probabilidade  $\beta$  então
9:          $S(t) \leftarrow S(t) - \{j\}$ 
10:         $I(t) \leftarrow I(t) \cup \{j\}$  // atualização dos estados
11:      fim se
12:      se  $j \notin S(t - 1)$  &  $i$  se torna contido com probabilidade  $\mu$  então
13:         $I(t) \leftarrow I(t) - \{i\}$ 
14:         $R(t) \leftarrow R(t) \cup \{i\}$ 
15:      fim se
16:    fim para
17:  fim para
18: até  $\phi(t) = 0$ 
19: retorna  $\varphi(t)$ 

```

infecção e a recuperação não ocorrem na mesma janela ou passo de tempo. A mudança de estado bem-sucedida (para ser informante ou para ser contido) ocorre só na seguinte iteração.

A configuração inicial para as simulações é $\psi(0) = 1 - 1/N$, $\phi(0) = 1/N$ e $\varphi(0) = 0$. Cada simulação começa com uma seleção uniforme de vértices conforme definido. Em cada passo de tempo, todos os informantes uniformemente selecionam um de seus vizinhos e tentam infectá-lo com probabilidade β , ou param a propagação com probabilidade μ . As simulações são executadas até o final do processo de difusão, quando $\phi(\infty) = 0$.

Adotamos aqui a abordagem do modelo de rumor de [Maki e Thompson \(1973\)](#) (MT) como a dinâmica de propagação. No processo MT sempre que um propagador ativo i contata um vizinho j que esteja inativo (ignorante), este último se tornará um propagador ativo com uma probabilidade fixa β . Caso contrário, se j está ciente sobre o rumor, ou seja j é um propagador ativo ou um contido, o vértice i se transformará em um propagador inativo (contido) com probabilidade μ . O comportamento quando o propagador para de ser ativo pode ser entendido dado que a informação é muito conhecida (ao contatar outros propagadores) ou sem novidade (ao contatar contidos). As regras gerais de contato para MT estão descritas na Equação 2.57.

O algoritmo para simular a propagação do rumor recebe como entrada a rede G e o conjunto de estados iniciais dos vértices $S(0), I(0), R(0)$. Os conjuntos de estados de vértices podem ser abordados como lista estruturadas de nós em um tempo específico. Para fins ilustrativos, os

passos (8 - 9) e (12 - 13) são o mesmo procedimento que atualiza o estado do vértice, mas podem ser generalizados em uma função separada.

A complexidade do algoritmo MC é determinada pelos seguintes custos: (i) número máximo de passos T , no qual $\phi(T) = 0$; (ii) número de iterações no laço principal da propagação (passi 4) que é $|S(t)|$ e podemos aproximá-lo a $N/2$; (iii) custo computacional de copiar e atualizar os vetores de estados, que podemos atribuir um valor constante C_1 ; e (iv) custo de contatar os vizinhos dos vértices, que no caso de contatar a todos os vizinhos, a média da rede é equivalente a $\langle k \rangle$. Assim, a complexidade de tempo do algoritmo de simulação é $\mathcal{O}(T * N/2 * C_1 * (2\langle k \rangle + 1)) \equiv \mathcal{O}(T * N * \langle k \rangle)$.

Para implementar a simulação do modelo SIR de rumor (Maki e Thompson), existem três principais abordagens de comportamento dos propagadores ([BORGES-HOLTHOEFER et al., 2012](#); [MELONI et al., 2012](#)). As abordagens são o processo reativo (RP), processo truncado (TP) e o processo por contato (CP). Essas abordagens estão relacionadas com detalhes de implementação e o comportamento que os propagadores têm no momento de difundir a informação. A seguir são descritas as características de cada uma das abordagens.

- **Processo reativo:** em cada iteração, cada vértice informante tenta informar todos seus vizinhos suscetíveis ou ignorantes e só depois disso, avalia se vai se tornar contido na seguinte iteração com seus vizinhos contidos ou informantes.
- **Processo truncado:** em cada iteração, cada vértice informante I_j seleciona aleatoriamente um vizinho e avalia a regra de transição de estado que corresponda. A seleção continua até que todos os vizinhos tenham sido contatados ou interrompida caso o informante I_j se torne contido, o que acontecer primeiro.
- **Processo de contato:** em cada iteração e para cada vértice informante I_j escolhe-se aleatoriamente um único vizinho. Depois, é avaliada a regra de transição de estado que corresponda e termina a vez desse informante passando para o seguinte da lista.

No entanto, uma análise clara sobre o impacto desses comportamentos na propagação ainda não tinha sido abordada na literatura (ver discussão na Seção [5.6.1](#)).

3.3 Redes complexas consideradas

Muitos sistemas complexos do mundo real podem ser representados como grafos ou redes. Uma classificação bastante conhecida é a proposta por Newman ([NEWMAN, 2003](#)), apresentada a seguir:

- **Redes sociais:** Uma rede social representa a conexão entre atores (vértices) mediante uma relação explícita. Um dos trabalhos mais conhecidos relacionado ao estudo de redes sociais

é o trabalho de Milgram e a famosa frase de “sete graus de separação”. Milgram (1967) afirmava que qualquer pessoa nos Estados Unidos estava a sete conhecidos de distância de qualquer outra pessoa do mesmo país. Existem vários exemplos de redes sociais, como as redes de colaboradores no caso de atores que tem atuado juntos em algum filme ou autores e coautores de artigos acadêmicos, redes de amizade, de trabalho, entre outras.

- **Redes de informação:** O exemplo clássico das redes de informação é uma rede de citações de artigos acadêmicos. Estas são redes dirigidas, pois um artigo cita a outro e a estrutura da rede de citações se assemelha com a estrutura da informação armazenada nos vértices. As redes de citações são acíclicas, pois um artigo não pode citar outro que ainda não foi escrito. Desse modo, os vértices da rede seguem um fluxo cronológico e portanto não existem loops. Outro exemplo de rede de informação é a *World Wide Web* (Web). Nessa rede, as páginas representam os vértices e os links as respectivas conexões. A diferença da rede de citações, a Web não apresenta restrições de conexão das páginas fazendo com que apareçam *loops* e que a rede seja cíclica.
- **Redes tecnológicas:** Um bom exemplo de redes tecnológicas são as redes de força elétrica, as quais transmitem a energia a todas as cidades de um país. São redes desenhadas tipicamente para a distribuição de algum bem ou recurso. Outros exemplos dessas redes são a rede de rodovias, a rede de tráfego aéreo e aeroportos, rede de conexão telefônica, e inclusive, a própria Internet.
- **Redes biológicas:** As redes biológicas são todas aquelas que mapeiam um sistema de seres vivos ou orgânico. Por exemplo, as redes ou cadeias alimentar de um ecossistema, onde alguns são a presa e outros o caçador. As redes biológicas abrangem desde a análise de redes de proteínas e redes genéticas, passando por redes corticóides e neuronais, até as próprias redes que representam grandes ecossistemas.

As seguintes redes complexas foram consideradas nos trabalhos de pesquisa abordados nesta tese:

1. *Google+* (MCAULEY; LESKOVEC, 2012; KUNEGIS, 2017) é uma rede dirigida sem peso nas arestas que contém links usuário-usuário do *Google Plus*. As arestas direcionadas denotam que um usuário adicionou o outro em seus círculos. O maior componente possui $N = 23613$ vértices e $M = 39183$ arestas, com o grau do maior *hub* igual a 2761.
2. *netscience* (NEWMAN, 2003) é uma rede sem peso e não dirigida de colaboração de cientistas que trabalham na área de redes complexas¹. Os dados foram compilados a partir das referências bibliográficas dos artigos de revisão da literatura de Newman (2003) e

¹ Dados da rede compilados por Mark Newman, disponível em <<http://www-personal.umich.edu/~mejn/netdata/>>

[Boccaletti et al. \(2006\)](#). Os vértices representam os pesquisadores cujos nomes aparecem nos artigos referenciados nesses dois trabalhos e as arestas conectam os autores que colaboraram nos artigos referenciados. O maior componente da rede possui 360 vértices e 2742 arestas, onde o maior *hub* possui grau igual a 34.

3. *astrophysics* ([NEWMAN, 2001](#)) é uma rede de colaboração de cientistas de trabalhos prévios de astrofísica, entre 1 de janeiro de 1995 até 31 de dezembro de 1999, arquivados em arxiv². A rede é ponderada e não dirigida e conta com 1029 componentes desconexos. O maior componente tem 15852 vértices e 119896 arestas e o maior *hub* do componente tem grau igual a 360.
4. *advogato* ([KUNEGIS, 2017](#); [MASSA; SALVETTI; TOMASONI, 2009](#)) é a rede de uma comunidade on-line para desenvolvedores de software livre lançada em 1999. Os vértices são usuários de *advogato* e as arestas direcionadas representam relações de confiança. O maior componente possui $N = 5054$ e $M = 39375$, com o grau do maior *hub* igual a 807.
5. *hamsterster* ([KUNEGIS, 2017](#)) é uma rede não direcionada e não ponderada baseada nos dados do site hamsterster.com. Os vértices são os usuários do sistema e as arestas representam a relação de amizade e família entre os usuários. A rede possui 2426 vértices, com 16098 arestas e maior *hub* com grau 273.
6. *internet* ([NEWMAN, 2013](#)) é uma rede que representa uma fração da estrutura simétrica da Internet, com os dados compilados a partir da reconstrução de tabelas BGP publicadas no projeto routerviews.org. A rede apresenta $N = 22963$ vértices, $M = 48436$ arestas e o maior *hub* da rede tem grau 2390, constituindo-se em apenas um componente.
7. *powergrid* ([WATTS; STROGATZ, 1998b](#)) trata-se de uma rede não direcionada que representa a topologia da rede elétrica do oeste dos Estados Unidos. A rede está disponível no site pessoal do autor³. É uma rede conexa com $N = 4941$ vértices e $M = 6593$ arestas e o vértice com maior grau tem valor 19. A rede apresenta uma estrutura mais regular, com poucas conexões entre os vértices, os quais estão mais distantes entre si (o diâmetro da rede é $\ell_{max} = 46$).
8. *email* ([GUIMERA et al., 2003](#)) representa uma rede social de intercâmbio de informação (correios eletrônicos) entre membros da *Univeristy Rovira i Virgili* (Tarragona). É uma rede direcionada e conexa que apresenta $N = 1133$ vértices, $M = 5451$ arestas e o maior *hub* da rede tem grau 71.
9. *caida* ([LESKOVEC; KLEINBERG; FALOUTSOS, 2005](#)) é uma rede não direcionada onde os vértices são sistemas autônomo da Internet com comunicação entre eles, coletados

² www.arxiv.org

³ <http://cdg.columbia.edu/cdg/datasets>

em 2007 a partir do projeto CAIDA. Seu maior componente possui $N = 26475$ vértices e $M = 53381$ arestas, com o grau do maior *hub* igual a 2628.

10. *PGP* (NÁ *et al.*, 2004) é a rede do maior componente dos usuários do algoritmo Pretty-Good-Privacy para o intercâmbio seguro de informações. O tamanho da rede é de 10680 vértices com 24317 arestas e maior *hub* com 205 conexões.

Assim, as bases *PGP*, *hamsterster*, *advogato* e *Google+* representam redes sociais. As redes de informação estão representadas pelas bases *email*, *astrophysics* e *netscience*. As bases *internet*, *powergrid* e *caida* são redes tecnológicas. Não foram consideradas as redes biológicas nas análises realizadas. Observamos que redes de um tipo também podem pertencer a outro, como a rede *email*, ao ter uma natureza do tipo rede social e ao mesmo tempo ser uma rede de intercâmbio de informação, ou a rede *internet* que pode ser considerada como uma rede de informação. Por fim, assumimos todas as redes como não direcionadas e sem peso nas arestas e adotamos sempre o seu componente principal.

MODELO HETEROGÊNEO DE PROPAGAÇÃO DE RUMORES

Modelos de propagação em redes complexas tem servido para prever e controlar o espalhamento de doenças ou rumores. A maioria dos modelos consideram probabilidades de propagação idênticas para todos os indivíduos. No entanto, esse pressuposto homogêneo é uma propriedade raramente encontrada na natureza e em redes sociais e tecnológicas. Neste capítulo apresentamos um modelo de propagação com probabilidades de transmissão heterogêneas. Analisamos o impacto da topologia de rede considerando a probabilidade de propagação proporcional a seis medidas de centralidade, avaliando em redes artificiais e do mundo real. Os resultados mostram que as probabilidades de transmissão correlacionadas ao PageRank, grau e centralidade de intermediação (*betweenness centrality*) melhoram a propagação nas redes sem escala em comparação com a abordagem homogênea. Mostramos também que a propagação pode ser controlada através de ajustes nas probabilidades de transmissão dos principais propagadores influentes. Os resultados contribuem para a compreensão da interação entre as dinâmicas de propagação e a transmissão heterogênea em redes sociais. O texto a seguir baseia-se nos resultados da publicação em (VEGA-OLIVEROS; COSTA; RODRIGUES, 2017b).

O capítulo está estruturado como segue: na Seção 4.1 apresentamos a motivação e lacuna na área relacionada com a heterogeneidade na transmissão de informação. A Seção 4.2 traz as abordagens e trabalhos relacionadas com o tema pesquisado. Na Seção 4.3 apresentamos a proposta do modelo heterogêneo para propagação de rumores em redes complexas. As redes complexas consideradas e a metodologia empregada neste trabalho são apresentadas nas Seção 4.4 e Seção 4.5 respectivamente. Os resultados das simulações e do modelo teórico são apresentados na Seção 4.6. A Seção 4.7 apresenta dois cenários de estudo que exemplificam os resultados obtidos pela proposta. Finalmente, as considerações finais do capítulo estão na Seção 4.8.

4.1 Motivação

Muitos modelos têm sido desenvolvidos para prever, controlar e compreender a propagação de ideias, conceitos e notícias por meio de redes (PASTOR-SATORRAS; VESPIGNANI, 2001; HETHCOTE, 2006; CASTELLANO; FORTUNATO; LORETO, 2009; FUNK; SALATHÉ; JANSEN, 2010). A modelagem de processos de difusão em redes sociais tem sido relatada em diferentes contextos, tais como a identificação de propagadores influentes (KITSACK *et al.*, 2010; BORGE-HOLTHOEFER *et al.*, 2012; BAE; KIM, 2014), rumores ou disseminação de conteúdos (DALEY; KENDALL, 1964; MAKI; THOMPSON, 1973; MORENO; NEKOVEE; PACHECO, 2004; BAKSHY *et al.*, 2012; HA *et al.*, 2015) e o surgimento de protestos políticos (GONZÁLEZ-BAILÓN; Borge-Holthoefer, Javier Rivero; MORENO, 2011), entre outros.

Os primeiros modelos de propagação de rumores consideraram apenas populações homogêneas misturadas, isto é, interações descritas como grafos completos (KEELING; ROHANI, 2008). No entanto, essa suposição ignora o cenário real em que os indivíduos interagem com uma pequena fração da população. Modelos mais precisos foram relatados desde o estabelecimento das redes complexas (BARABÁSI; ALBERT, 1999; NEWMAN, 2003; BOCCALETTI *et al.*, 2006). A representação de sistemas complexos como redes possibilitou a inclusão de estruturas heterogêneas nos modelos tradicionais, o que influencia fortemente a disseminação da informação (PASTOR-SATORRAS; VESPIGNANI, 2001; MORENO; NEKOVEE; PACHECO, 2004).

Embora a estrutura da rede forneça uma modelagem mais realista para a propagação de informação, a maioria das abordagens consideram que a transmissão é homogênea (DALEY; KENDALL, 1964; MAKI; THOMPSON, 1973; CASTELLANO; FORTUNATO; LORETO, 2009; PASTOR-SATORRAS *et al.*, 2015). No entanto, o suposto de homogeneidade não é uma propriedade observada nas redes reais como as sociais e tecnológicas. Por exemplo, doenças como o sarampo, caxumba ou varicela são mais comumente desenvolvidas durante a infância e a transmissão entre as crianças é muito mais provável de ocorrer do que entre adultos (KEELING; ROHANI, 2008). Além disso, os indivíduos nas redes sociais não propagam informações com a mesma intensidade — celebridades e líderes políticos são propagadores mais influentes do que as pessoas comuns. Para um modelo de propagação mais preciso e real, as abordagens com transmissão heterogênea devem ser levadas em consideração.

4.2 Trabalhos relacionados

A maioria das abordagens desenvolvidas têm negligenciado o comportamento heterogêneo dos indivíduos nos processos de propagação em redes. Em epidemias, o conceito da abordagem *dealing risks* (KEELING; ROHANI, 2008) é algumas vezes adotado, o qual divide a população em classes com características comportamentais semelhantes (FUNK; SALATHÉ; JANSEN, 2010). Este modelo pressupõe uma população com mistura homogênea onde os mem-

bro da mesma classe têm risco equivalente de contrair e transmitir a epidemia (KEELING; ROHANI, 2008). Poderia-se estender os modelos *dealing risks* à estrutura de rede para gerar um modelo de propagação heterogêneo que lide com classes de vértices. No entanto, incluir a heterogeneidade local neste modelo aumenta o número de equações em comparação com o cenário homogêneo. Por exemplo, assumindo apenas as classes de alto e baixo risco de contágio para os modelos *SIR*, o número final de equações dobraria, aumentando também o custo computacional.

Por outro lado, Meloni, Arenas e Moreno (2009) estudaram a propagação de epidemias com transmissão heterogênea considerando um fluxo de tráfego de pacotes em redes, representando-o como uma maior probabilidade de receber a doença. Os autores adotaram o modelo suscetível-infectado-susceptível (*SIS*) e as capacidades de processamento e transmissão da informações foram heterogeneamente distribuídas de acordo com a medida de centralidade de intermediação. Os autores verificaram que o valor do limiar epidêmico em redes sem escala depende diretamente do primeiro e segundo momentos da distribuição de centralidade de intermediação. No entanto, a influência que outras medidas de centralidade, como grau ou PageRank, podem ter sobre a dinâmica de difusão não foi abordada (MELONI; ARENAS; MORENO, 2009).

De fato, estudos sobre a propagação de rumores em redes com transmissão heterogênea não têm sido abordados (ver por exemplo (CASTELLANO; FORTUNATO; LORETO, 2009; PASTOR-SATORRAS *et al.*, 2015)). Alguns estudos analisaram a heterogeneidade individual para redes espaço-temporais. Em Loglisci, Ceci e Malerba (2015), os autores distinguem entre diferentes tipos de atributos para vértices e arestas. A representação foi baseada em um programa de lógica indutiva que permite fazer raciocínio de primeira ordem. Assim, é possível, mediante predicados lógicos, inferir relações de elementos e encontrar padrões em dados. Embora essa abordagem considere características heterogêneas, o objetivo principal foi a mineração de dados de rede e não um modelo de representação baseado em redes ou a análise de processo de propagação. O comportamento heterogêneo para os indivíduos é particularmente importante para o processo de propagação e pode ser considerado para obter uma modelagem mais completa.

4.3 Proposta do modelo heterogêneo

Nós propomos um modelo de rumor em tempo discreto baseado em cadeias de Markov com probabilidades heterogêneas. A motivação para este modelo reside no fato de que os padrões de atividade na dinâmica social são em sua maioria heterogêneos e nem todos os indivíduos têm a mesma capacidade de convencimento (exemplo (MELONI; ARENAS; MORENO, 2009)). As probabilidades que um vértice x seja ignorante ($S_x(t)$), informante ($I_x(t)$) ou contido ($R_x(t)$) satisfaz $S_x(t) + I_x(t) + R_x(t) = 1$ para todo tempo t . A disseminação da informação pode ser descrita pelas regras gerais de rumor apresentadas na Equação 2.57. Neste caso, a probabilidade

de um vértice x estar em um dos estados possíveis no momento $(t + 1)$ é calculada por

$$S_x(t + 1) = S_x(t) \bar{i}_x(t), \quad (4.1)$$

$$I_x(t + 1) = I_x(t) \bar{r}_x(t) + S_x(t) [1 - \bar{i}_x(t)], \quad (4.2)$$

$$R_x(t + 1) = R_x(t) + I_x(t) [1 - \bar{r}_x(t)], \quad (4.3)$$

onde $\bar{i}_x(t)$ é a probabilidade de que o vértice x não se torne um propagador por algum dos seus vizinhos no instante de tempo t . De modo similar, $\bar{r}_x(t)$ é a probabilidade de que x não se torne contido no tempo t . Na Equação 4.2, o vértice x é um informante no instante $(t + 1)$ se: (i) x era um propagador no passo anterior (com probabilidade $I_x(t)$) e não se tornou contido durante o tempo atual, que ocorre com a probabilidade $\bar{r}_x(t)$; ou (ii) x era um ignorante no tempo t (com probabilidade $S_x(t)$) e foi informado por qualquer um dos seus vizinhos de acordo com a probabilidade $[1 - \bar{i}_x(t)]$. A probabilidade do vértice x ser um contido no instante $(t + 1)$, Equação 4.3, depende das seguintes condições: (i) se x era um informante que tornou-se um contido por qualquer um de seus vizinhos informados, o qual acontece com probabilidade $I_x(t) [1 - \bar{r}_x(t)]$; ou (ii) x era um contido no tempo anterior (com probabilidade $R_x(t)$) e permaneceu nesse estado no seguinte tempo $(t + 1)$.

A probabilidade de que o vértice x não se torne um informante ou contido no instante t , isto é, a probabilidade de que ele não realize uma transição para outro estado é dada por

$$\bar{i}_x(t) = \prod_{y=1}^N \left[1 - \beta_y \frac{A_{yx}}{k_y} I_y(t) \right], \quad (4.4)$$

$$\bar{r}_x(t) = 1 - \sum_{y=1}^N \mu_x \frac{A_{xy}}{k_x} [I_y(t) + R_y(t)]. \quad (4.5)$$

A probabilidade de que o vértice x não se torne um informante depende de se algum de seus vizinhos não o tenha infectado (Equação 4.4). O vértice x pode ser infectado no tempo t por algum dos seus vizinhos informantes y com probabilidade de propagação β_y . Desse modo, a probabilidade de x não se torna um informante por y é $\left[1 - \beta_y \left(\frac{A_{yx}}{k_y} \right) I_y(t) \right]$. O produtório em $\bar{i}_x(t)$ deve-se ao pressuposto de que cada vizinho entra em contato com x de modo independente.

Por outro lado, um propagador x pode se tornar contido ao entrar em contato com algum vizinho informado (Equação 4.5). Se o vértice x , com probabilidade de recuperação μ_x , contata um vizinho que é informado, então x se torna um contido de acordo com essa probabilidade. A probabilidade máxima de y ser informado ou ciente do rumor ocorre quando $S_y(t) = 0$, dado que $S_y(t) + I_y(t) + R_y(t) = 1$. Um vértice informado é selecionado aleatoriamente de acordo com o conjunto de vizinhos informados conectados a x .

Nesta formulação assumimos que a infecção e a recuperação não ocorrem no mesmo passo de tempo. Além disso, adotamos apenas um contato aleatório uniforme por unidade de tempo, entre um propagador e um vizinho, conhecido como processo de contato (CP). Assim, a probabilidade de selecionar um vizinho de x é $A_{xy}/(k_x)$.

A propagação heterogênea é incorporada no modelo introduzindo um vetor característico $\vec{\beta} = \{\beta_1, \beta_2, \dots, \beta_N\}$, cujos elementos representam a probabilidade de propagação para cada vértice. Da mesma forma, o vetor $\vec{\mu} = \{\mu_1, \mu_2, \dots, \mu_N\}$ retorna as probabilidades de recuperação. O modelo de rumor clássico com propagação homogênea pode ser recuperado fazendo $\beta_x = \beta$ e $\mu_x = \mu, \forall x$, onde β e μ são constantes.

Na Equação 4.2, substituindo $I_x(t)$ e depois de algumas operações algébricas, a probabilidade do vértice x ser infectado é dada por

$$I_x(t+1) = S_x(t) [1 - (\bar{r}_x(t) + \bar{i}_x(t))] + (1 - R_x(t))\bar{r}_x(t). \quad (4.6)$$

Avaliando a Equação 4.1 em um intervalo definido $[0, t]$, em termos dos valores de progressão dos vértices, obtemos

$$S_x(t) = S_x(0) \prod_{t'=0}^t \bar{i}_x(t'). \quad (4.7)$$

A resolução do modelo é obtida resolvendo o sistema para $x = 1, \dots, N$ e $t \rightarrow \infty$,

$$\begin{aligned} S_x &= S_x \bar{i}_x, \\ I_x &= -S_x [\bar{r}_x] + (1 - R_x) \bar{r}_x, \\ R_x &= R_x + I_x [1 - \bar{r}_x], \end{aligned} \quad (4.8)$$

onde a classe dos contidos (R) é o estado absorvente de todos os propagadores. Assim, quando $t \rightarrow \infty$, temos apenas ignorantes ou contidos na rede.

A densidade final de contidos ($\varphi(\infty)$) é obtida pela Equação 4.6 no limite quando $t \rightarrow \infty$. Após a inclusão do resultado da Equação 4.7 e de algumas operações, obtemos a densidade de indivíduos informados na rede em um instante de tempo t ,

$$\zeta(t) \approx 1 - \frac{1}{N} \sum_{x=1}^N \prod_{t'=0}^t \bar{i}_x(t'). \quad (4.9)$$

Conseqüentemente, no final do processo de propagação temos que $\zeta_\infty = \varphi(\infty)$.

Os resultados teóricos das abordagens homogêneas e heterogêneas são comparados com simulações de Monte Carlos (MC), em micro e macro escalas. A simulação considera tempo discreto e cada simulação começa com um conjunto aleatório de propagadores iniciais. Na abordagem clássica (homogênea), os propagadores infectam seus vizinhos ignorantes com a mesma probabilidade β e se recuperam com a probabilidade μ . A taxa de propagação é definida

Figura 12 – Caso homogêneo. Propagação de rumor para uma rede Barabási-Albert (BA) com tamanho de $N = 10^4$ vértices: (a) densidade final de contidos (vermelho = 1, azul = 0) para diferentes taxas de propagação (linhas brancas tracejadas). Valores iguais de λ produzem a mesma densidade esperada de contidos; (b) comparação dos resultados teóricos (linha contínua) com simulações de Monte Carlo (MC) (símbolos). As figuras internas mostram a probabilidade de que o vértice x ($x = 1, \dots, N$) seja um contido (R_x , pontos vermelhos) ou um ignorante (S_x , pontos verdes) para simulações de MC e resultados teóricos. Cada simulação é uma média de mais de 500 realizações.

Fonte: Adaptada de [Vega-Oliveros, Costa e Rodrigues \(2017b\)](#).

por $\lambda = \beta/\mu$. Na Figura 12a mostramos a densidade de contidos de acordo com diferentes valores de λ . Podemos ver que nenhuma transição de fase é observada na propagação de rumores, ao contrário da propagação da epidemias ([PASTOR-SATORRAS *et al.*, 2015](#)).

A comparação da precisão entre a simulação de Monte Carlo (MC) e a solução numérica é apresentada na Figura 12b para a propagação homogênea. As simulações começam com uma seleção uniforme de $\phi(0) = 0.01$ propagadores iniciais. Sem perda de generalidade, assumimos $\mu = 1$ e valores diferentes de β . As figuras internas mostram que os resultados teóricos capturam a dinâmica microscópica das simulações de MC.

A análise teórica também é comparada com as simulações de MC para o caso da propagação heterogênea, como mostrado na Figura 13. A probabilidade de propagação β_x é assumida como sendo proporcional ao valor de certa medida de centralidade para o vértice. Podemos ver que as soluções analíticas concordam muito bem com os resultados da simulação de MC. Assim, os resultados teóricos são precisos para ambos os casos heterogêneos ou homogêneos.

4.4 Redes consideradas

Exploramos a propagação heterogênea em redes artificiais e do mundo real. Para redes artificiais, consideramos as redes aleatórias de Erdős-Rényi (ER) ([ERDÖS; RÉNYI, 1959](#)), de pequeno mundo de Watts-Strogatz (WS) ([WATTS; STROGATZ, 1998a](#)), sem escala Barabási-

Figura 13 – Caso heterogêneo. Densidade final de contidos em um rede Barabási-Albert (BA) de tamanho $N = 10^4$. Os símbolos correspondem as simulações de Monte Carlo (MC) enquanto que as linhas contínuas são soluções da análise teórica. A probabilidade de propagação β_x é proporcional às medidas de centralidade: (a) grau, (b) intermediação e (c) PageRank. As figuras internas mostram a probabilidade de que o vértice x ($x = 1, \dots, N$) seja um contido (R_x , pontos vermelhos) ou um ignorante (S_x , pontos verdes) para as simulações de MC e previsões teóricas. Cada simulação é uma média de 500 realizações.

Fonte: Adaptada de [Vega-Oliveros, Costa e Rodrigues \(2017b\)](#).

Tabela 1 – Propriedades estruturais das redes complexas analisadas.

	Rede	N	$\langle k \rangle$	$\langle \ell \rangle$	ρ	$\langle B_x \rangle$	$\langle C\ell_x \rangle$	$\langle X_x \rangle$	$\langle Kc_x \rangle$	$\langle \pi_x \rangle$
Artificial	ER	1000	12	3.04	1.186×10^{-3}	1.018×10^3	3.30×10^{-4}	0.030	7.81	1.0×10^{-3}
	WS	1000	12	3.59	-1.665×10^{-2}	1.684×10^3	2.29×10^{-4}	0.031	5.98	1.0×10^{-3}
	BA	1000	12	2.84	-3.221×10^{-2}	9.204×10^2	3.54×10^{-4}	0.022	6.00	1.0×10^{-3}
	SSF	1000	12	4.08	-1.963×10^{-1}	1.536×10^3	2.48×10^{-4}	0.022	6.08	1.0×10^{-3}
Real	<i>netscience</i>	379	4.82	6.04	-0.082	9.529×10^2	4.51×10^{-4}	0.016	3.47	2.64×10^{-3}
	<i>email</i>	1133	9.62	3.60	0.078	1.475×10^3	2.49×10^{-4}	0.017	5.35	8.83×10^{-4}
	<i>hamsterster</i>	2000	16.1	3.59	0.022	2.588×10^3	1.43×10^{-4}	0.011	9.29	5.00×10^{-4}
	<i>google+</i>	23613	3.32	4.03	-0.388	3.581×10^4	1.10×10^{-5}	2.30×10^{-3}	1.67	4.21×10^{-5}

Albert (BA) ([BARABÁSI; ALBERT, 1999](#)) e redes espaciais de Barthélemy (SSF) ([BARTHÉLEMY, 2003](#)).

Consideramos dados de redes reais de colaboração entre pesquisadores, de troca de emails e de amizade ou contato social: (i) rede *netscience*, que é uma rede de colaboração de cientistas; (ii) *email*, que é a rede que representa a troca de emails entre membros de uma universidade; (iii) *hamsterster*, a rede de contato de amizade entre usuários de um portal de donos de mascotes; e (iv) *Google+*, uma parte da rede de usuários da rede social Google Plus. A descrição dessas redes se encontra na Seção 3.3. As medidas topológicas das redes consideradas são apresentadas na Tabela 1. Apresentamos o grau médio $\langle k \rangle$, a média dos caminhos mínimos $\langle \ell \rangle$, a média do grau de intermediação $\langle B_x \rangle$, a média da centralidade de proximidade $\langle C\ell_x \rangle$, a média da centralidade de autovetor $\langle X_x \rangle$, da centralidade de K -core $\langle Kc_x \rangle$, do PageRank $\langle \pi_x \rangle$, e coeficiente de assortatividade ρ . Podemos ver que a maioria das redes são esparsas e têm um comprimento médio de caminhos mínimos curto.

4.5 Metodologia

Correlacionamos as probabilidades de transmissão com às medidas de centralidade justamente porque se espera que quanto mais central seja um sujeito, maior seu impacto na difusão de informação. Realizamos então as simulações de Monte Carlo para analisar a influência da distribuição de probabilidades heterogênea na dinâmica de propagação. Empregamos a capacidade de propagação de informação do vértice x , a qual definimos como a fração final de indivíduos informados (contidos) obtida quando a simulação começa em x , ou seja, quando x é selecionado como o único propagador inicial. Os resultados são obtidos com a média de 60 simulações.

Também, para o caso macroscópico consideramos a média das capacidades de propagação de todos os vértices da rede e como regra de comportamento dos propagadores o processo de contato (CP). Para fins de comparação do impacto das probabilidades heterogêneas com o caso clássico, consideramos a seguinte função de permutação:

$$\beta_x = p\langle\tilde{\theta}\rangle + (1 - p)\tilde{\theta}_x, \text{ com } 0 \leq p \leq 1, \quad (4.10)$$

para assim definir a probabilidade de propagação de informação de cada vértice x . Desta forma, podemos controlar o nível de heterogeneidade na propagação, definindo o parâmetro p em $[0, 1]$. A propagação heterogênea ocorre quando $p < 1$, enquanto que $p = 1$ recupera o caso homogêneo clássico, no qual todos os vértices têm a mesma capacidade de transmissão da informação. Além disso, uma vez que as medidas de centralidade são definidas em diferentes escalas, consideramos a versão normalizada de cada medida de centralidade, ou seja, $\tilde{\theta}_x = \theta_x / \max_{x=1,\dots,N}(\theta_x)$. Assim, $0 \leq \tilde{\theta}_x \leq 1$, $x = 1, 2, \dots, N$. Finalmente, mantemos a média das probabilidades de propagação constante, $\langle\tilde{\theta}\rangle = \beta$, independente do valor de p . Esta suposição é necessária para realizar uma comparação justa entre os casos homogêneos e heterogêneos. Também, para a realização das simulações, assumimos a probabilidade de recuperação $\mu = \langle\tilde{\theta}_x\rangle / \sqrt{2}$, sem perda de generalidade nos resultados.

4.6 Resultados

4.6.1 Redes artificiais

Para verificar como a estrutura da rede influencia a propagação do rumor, consideramos redes artificiais geradas pelo modelo aleatório de Erdős-Rényi (ER), modelo de redes pequeno mundo de Watts-Strogatz (WS), modelo de redes sem escala Barabási-Albert (BA) e o modelo de redes espaciais sem escala de Barthélemy (SSF) (BARTHÉLEMY, 2003). As redes foram geradas considerando o mesmo número de vértices e arestas, uma vez que estas quantidades estão relacionadas com o processo de propagação (por exemplo, a propagação é mais rápida em redes mais densas, ver (PASTOR-SATORRAS *et al.*, 2015)). Podemos ver na Figura 14 que a propagação heterogênea não afeta significativamente a fração final de contidos nas redes Poisson

e de pequeno mundo. Esses resultados são esperados, uma vez que estas redes apresentam vértices com níveis de centralidade muito similares, devido à homogeneidade de sua organização. No entanto, em redes sem escala, a propagação heterogênea afeta a fração final de contidos quando a probabilidade de propagação β_x é proporcional ao grau, centralidade de intermediação, PageRank e centralidade de autovetor. No caso da centralidade de intermediação, o número máximo de contidos é atingido quando $p \approx 0.5$. Assim, neste caso, um nível parcial de centralidade é a melhor configuração para atingir o maior número de usuários. Para as outras centralidades, isto é, grau, PageRank e centralidade de autovetor, a propagação completamente heterogênea fornece a melhor capacidade de propagação do rumor. Em relação à centralidade de proximidade e a centralidade k -core, a fração final de contidos não é afetada pela propagação heterogênea. Isso ocorre porque essas medidas produzem valores de centralidade muito próximos (NEWMAN, 2010). Desta forma, os vértices apresentam probabilidades de propagação semelhantes, mesmo no caso completamente heterogêneo. Comparando todas as medidas, a probabilidade de propagação proporcional ao grau e PageRank rende as maiores frações de indivíduos informados.

Como observamos que a correlação entre as medidas de centralidade e a probabilidade de propagação aumenta a capacidade de propagação, é importante verificar o impacto dessa correlação. Assim, atribuímos a probabilidade de propagação para cada vértices estritamente proporcional às medidas de centralidade, ou seja, consideramos o caso completamente heterogêneo ($p = 0$). Em seguida, realizamos uma troca aleatória nessas probabilidades para diminuir a correlação entre a estrutura e a dinâmica. Ou seja, selecionamos uniformemente pares de vértices e trocamos suas probabilidades de propagação. Foram consideradas diferentes frações de pares de vértices para realizar as trocas, buscando verificar o impacto da correlação entre as propriedades estruturais dos nós e a capacidade de propagação. Na Figura 15, observamos que para as redes ER e WS, a troca das probabilidades não impacta na capacidade de difusão do rumor. Por outro lado, o embaralhamento afeta a fração final de contidos nas redes sem escala. Para as centralidades de intermediação, grau, PageRank e autovetor, a fração de indivíduos informados diminui quando a taxa de embaralhamento aumenta. No entanto, mesmo com 90% dos vértices com as probabilidades de propagação embaralhadas, a fração final de contidos é maior do que no caso homogêneo. Portanto, a correlação entre as medidas de centralidade do vértice e sua probabilidade de propagação da informação é importante para melhorar a disseminação em redes sem escala, como é o caso das redes sociais.

4.6.2 Redes sociais

As simulações também foram realizadas em dados de redes sociais. Podemos ver na Figura 30 que a fração final de contidos, em termos do nível de heterogeneidade na propagação de rumores, é semelhante para os modelos de redes artificiais sem escala. Para a centralidade de intermediação e centralidade de autovetor, a correlação parcial fornece a maior fração de indivíduos informados. Por outro lado, para o grau e o PageRank, a propagação completamente

Figura 14 – Densidade final de contidos de acordo com o nível de heterogeneidade definido pela função de interpolação (Eq. 4.10). As medidas de centralidade consideradas são centralidade de intermediação (BE), centralidade de proximidade (CL), grau (DG), centralidade de autovetor (EV), k -core (KC) e PageRank (PR). A propagação homogênea é dada para $p = 1$, enquanto que o caso completamente heterogêneo ocorre para $p = 0$. As simulações são realizadas considerando modelos de redes (a) Erdős-Rényi (ER), (b) Watts-Strogatz (WS), (c) Barabási-Albert (BA) e (d) espaciais sem escala de Barthélemy (SSF). A taxa de propagação $\lambda = \sqrt{2}$ foi adotada em todas as simulações.

Fonte: Adaptada de [Vega-Oliveros, Costa e Rodrigues \(2017b\)](#).

heterogênea produz a melhor propagação, atingindo as maiores frações de indivíduos informados. Para a centralidade k -core e de proximidade, a fração final de contidos não é fortemente afetada pelo nível de heterogeneidade.

Para as redes *netscience*, *email* e *hamster*, a correlação parcial entre o grau de intermediação (BE) e a probabilidade de propagação fornece a maior fração de contidos. Para a rede de *Google++*, a correlação com PageRank produz a maior fração de nós informados. Estes resultados são diferentes daqueles observados nos modelos de redes artificiais (ver Figura 14). Este fato é devido à organização altamente estruturada das redes sociais, que apresentam divisão modular e níveis consideráveis de triângulos e correlação de grau ([VEGA-OLIVEROS; BERTON, 2015](#)).

Figura 15 – Densidade final de contidos em função da proporção de trocas aleatórias da probabilidade de propagação β_x nas redes geradas pelos modelos: (a) Erdős-Rényi (ER), (b) Watts-Strogatz (WS), (c) Barabási-Albert (BA) e (d) espacias sem escala de Barthélemy (SSF). As medidas de centralidade empregadas são a centralidade de intermediação (*BE*), a centralidade de proximidade (*CL*), o grau (*DG*), centralidade de autovetor (*EV*), *k*-core (*KC*) e PageRank (*PR*).

Fonte: Adaptada de [Vega-Oliveros, Costa e Rodrigues \(2017b\)](#).

4.7 Cenários de estudo: melhora e controle da propagação

Para exemplificar o modelo, foram propostos dois cenários hipotéticos como motivação prática da aplicação do modelo proposto. A análise de comportamento heterogêneo foi feita na rede *Google+*, dado que as redes sociais exibem estruturas livres de escala ([NEWMAN, 2003](#); [BOCCALETTI et al., 2006](#); [HA et al., 2015](#)) e são caracterizadas por uma distribuição altamente heterogênea no número de conexões.

Figura 16 – Fração final de contidos de acordo com o nível de heterogeneidade definido pela função de interpolação (Eq. 4.10) nas redes sociais: (a) *netscience*, (b) *email*, (c) *hamsterster* e (d) *Google+*. As medidas de centralidade consideradas são: centralidade de intermediação (BE), centralidade de proximidade (CL), grau (DG), centralidade de autovetor (EV), *k*-core (KC) e PageRank (PR). A propagação homogênea é dada para $p = 1$, enquanto que o caso completamente heterogêneo ocorre para $p = 0$.

Fonte: Adaptada de Vega-Oliveros, Costa e Rodrigues (2017b).

4.7.1 Cenário de controle da propagação de desinformação

Nosso interesse inicial é no controle da difusão social de rumores, com aconteceu no caso real do cruzeiro *Carnival Magic* em outubro de 2014 ¹. Uma supervisora de laboratório que viajava no cruzeiro havia processado amostras clínicas de um homem que tinha sido diagnóstico com Ebola. O paciente morreu quatro dias antes da saída do navio. A supervisora e seu marido voluntariamente concordaram em ser colocados em quarentena em sua cabine para evitar a possível infecção e disseminação do vírus. Uma amostra de sangue foi retirada da mulher e os resultados do teste de Ebola foram negativos. As autoridades assumiram que não havia evidência de uma possível ameaça de saúde pública, para os passageiros do cruzeiro ou qualquer país fronteiriço. No entanto, a ansiedade e o medo dos passageiros, a falta de clareza e atraso na divulgação de informações oficiais pelas autoridades levou ao pânico e disseminação

¹ Jack Healy. (2014, 20 de outubro) Ebola Is Ruled Out After Cruise Ship Carrying Hospital Worker Returns to Texas. *The New York Times*. Retrieved from <http://www.nytimes.com>

de informações erradas. Este incidente mostra a importância de um melhor controle sobre a informação difundida pelas agências de saúde pública e governamentais através de alguma rede social durante eventos de emergência.

Em futuras situações, informações oficiais e esclarecedoras podem ser publicadas nos perfis de um pequeno grupo de indivíduos, selecionados de alguma rede social relevante, para assim esclarecer a situação. O objetivo desta estratégia é controlar os rumores assim que forem detectados, independente de se os indivíduos selecionados sejam parte ou não da disseminação da informação errônea.

4.7.2 *Cenário de melhora na propagação da informação*

O segundo cenário é um exemplo baseado no marketing viral. Uma empresa de moda está promovendo um novo produto através de seus canais regulais de publicidade, mas quer, de forma eficiente e de baixo custo, promover o produto para potenciais novos clientes. A empresa pretende selecionar um pequeno número de clientes co-participantes da blogosfera para que experimente gratuitamente o produto. Espera-se que tais clientes gostem do produto e comecem a postar boas recomendações para influenciar seus amigos e seguidores (BAKSHY *et al.*, 2012). Uma grande parte da blogosfera iria ouvir sobre o novo produto através do efeito "word-of-post". Esses co-participantes aliados são usuários que geralmente influenciam, intencionalmente ou não, a opinião de outros usuários quando publicam suas opiniões. Portanto, as empresas estão interessadas em descobrir esses influenciadores, por causa do impacto que exercem sobre as opiniões dos outros e como um potencial mecanismo para melhorar a promoção e venda de produtos e serviços. O problema é descobrir quais usuários selecionar, aqueles que sejam capazes de influenciar o maior número de pessoas visando o aumento na capacidade de difusão na rede.

4.7.3 *Resultados dos cenários de estudo*

Em ambos os cenários foram consideradas as seguintes opções para selecionar os vértices iniciais: *i*) selecionar os vértices com maior centralidade; *ii*) selecionar de modo aleatório. Definimos uma proporção $\tau = 0.5\%|V|$ de indivíduos alvo para ajudar a melhorar ou controlar a difusão da informação. Essa quantidade pequena foi motivada de acordo com as possíveis restrições e custos para realizar as ações ou campanhas nos cenários reais e capturar os indivíduos mais representativos segundo a medida de centralidade.

Para fins ilustrativos, chamamos a capacidade de propagação ao longo do tempo ($\zeta(t)^V$) como a média de indivíduos informados naquele respectivo instante t . A capacidade de propagação final (ζ_∞^V) é a fração de indivíduos informados no final da dinâmica, ou seja quando $t \rightarrow \infty$. Definimos também a capacidade de propagação do modelo clássico ou homogêneo ($\zeta_\infty^V C$), que corresponde ao modelo no qual todos os indivíduos exibem a mesma probabilidade de transmissão $\beta_x = \langle \beta \rangle$. As simulações começam da forma clássica antes de que uma proporção

Tabela 2 – Capacidade de propagação final de acordo com as centralidades

α (%)	Melhoria			Controle		
	$\zeta_{\infty}^V BE$	$\zeta_{\infty}^V DG$	$\zeta_{\infty}^V PR$	$\zeta_{\infty}^V BE$	$\zeta_{\infty}^V DG$	$\zeta_{\infty}^V PR$
1	1.164	1.215	1.224	0.796	0.713	0.697
3	1.154	1.198	1.209	0.813	0.741	0.725
6	1.125	1.160	1.168	0.850	0.798	0.787
10	1.052	1.073	1.082	0.934	0.915	0.907
12	1.014	1.012	1.014	0.983	0.979	0.977
15	1.000	1.000	1.000	1.000	1.000	0.995

de α indivíduos estejam cientes da informação. A razão disso é que pelo menos uma fração finita da rede ouvirá sobre as informações até que as ações correspondentes sejam executadas.

A medida de centralidade PageRank, em comparação com as outras métricas, forneceu os melhores resultados na seção anterior. No entanto, foi examinada qual é a centralidade mais adequada para os cenários. Primeiramente, a probabilidade de propagação β_x para o conjunto de τ indivíduos foi considerada proporcional ao grau (DG), grau de intermediação (BE) e PageRank (PR). Os indivíduos alvo (τ) foram selecionados de acordo com o ranking de tais centralidades em ordem decrescente. A capacidade de propagação final de cada medida é normalizada pela capacidade de propagação do modelo clássico, isto é, para o caso da centralidade de grau $\zeta_{\infty}^V DG = \zeta_{\infty}^{\tau(DG \downarrow)} / \zeta_{\infty}^V C$, com $\tau(DG \downarrow)$ significando a seleção em ordem decrescente dos vértices com maior valor de centralidade. De modo semelhante ocorre para as outras medidas de centralidade.

Na tabela 2 apresentamos a capacidade de propagação de cada medida de centralidade segundo o cenário de melhoria ou controle da difusão da informação. Os indivíduos-alvo com maior centralidade de PageRank $\zeta_{\infty}^V PR$ propagaram e controlaram melhor a difusão da informação. Se a proporção de indivíduos informados α aumenta, as ações tornam-se ineficazes. Consequentemente, para as seguintes análises, adotamos $\alpha = 1\%$, como o caso de detecção precoce do rumor, e consideramos apenas a centralidade de PageRank.

De acordo com a descrição dos cenários, todos os indivíduos-alvo receberão o mesmo incentivo. Portanto, eles são um subgrupo com a mesma probabilidade de propagação, mas que é diferente do restante da população. Isto é, $\beta_x^{\tau} = \langle \beta \rangle (1 \pm \Delta \beta)$ com $\langle \beta \rangle = 0.5$ e $\Delta \beta = 0.8$ como fator de escala nas simulações. Esse valor foi adotado porque existe uma pequena chance de que alguns dos indivíduos-alvo discordem em transmitir a informação. Como foi definido, os indivíduos-alvo foram selecionados de acordo com a classificação de maior centralidade de PageRank ($\tau(PR \downarrow)$). A desinformação é controlada pela diminuição da probabilidade de propagação do conjunto τ , sendo esta $\beta_x^{\tau} = 0.1$. Para o cenário de melhoria é aumentada para $\beta_x^{\tau} = 0.9$.

Também, foi considerado o efeito de selecionar aleatoriamente os indivíduos-alvo da rede inteira, mas não houve diferença nos resultados. Isso significa que a fração de $\tau = 0.5\%|V|$ foi muito baixa para gerar alguma variação nos resultados no caso aleatório. Verificamos que seria

Figura 17 – Diagramas de fase do comportamento da propagação ao longo do tempo na rede *Google++*. Considerou-se a abordagem de propagação clássica com β_x homogêneo. Foi melhorada ou controlada a capacidade de propagação selecionando um grupo de $\tau = 0.5\%$ vértices alvo e mudando suas probabilidades de propagação para um novo β_x^τ . A seleção dos τ vértices foi dos que tinha maior centralidade de PageRank. Também foi feita a seleção uniformemente aleatória de τ vértices para melhorar (Iprv) ou controlar (Ctrl) a difusão. (a) diagrama de fase dos propagadores ao longo do tempo normalizado pelo pico máximo da abordagem clássica; (b) diagrama de fase da capacidade de propagação final ao longo do tempo normalizado pela capacidade de propagação final da abordagem clássica.

Fonte: Adaptada de Vega-Oliveros, Costa e Rodrigues (2017b).

preciso um $\tau = 80\%|V|$ de indivíduos-alvo selecionados aleatoriamente para atingir resultados equivalentes aos obtidos pelo PageRank. Portanto, decidimos fazer a seleção aleatória dos indivíduos-alvo sobre a população que está informada.

A média de propagadores e indivíduos informados ao longo do tempo é mostrada na Figura 17. A curva “clássica” corresponde ao comportamento homogêneo. As curvas “*improved*” (melhorada) e “*controlled*” (controlada) são os resultados para os τ vértices com maior centralidade de PageRank. As outras curvas são os resultados de fazer a seleção aleatória para melhorar (Iprv) ou controlar (Ctrl) a propagação. As curvas foram normalizadas de acordo com o pico de propagadores (Figura 17a) e a capacidade de propagação final (Figura 17b) obtidas a partir da abordagem clássica. Com apenas $0.5\%|V|$ de vértices alvo, melhorou em 35% a capacidade de propagação e obteve-se uma redução de 60% em comparação com a abordagem clássica. A capacidade de propagação final dos τ indivíduos aleatórios não alcançou a metade do valor obtido pelos melhores vértices de PageRank, o que mostrou que o comportamento heterogêneo tem importantes implicações no processo de propagação para melhorar as aplicações em sistemas de recomendação, confiança e segurança.

4.8 Considerações finais

Neste capítulo foi apresentado um modelo de propagação de rumores com comportamento heterogêneo de propagação para redes sociais. Assumimos que as probabilidades de propagação de cada vértice são proporcionais a suas medidas de centralidade. Seis medidas de centralidade, a saber, grau, centralidade de intermediação, proximidade, k -core, PageRank e centralidade de autovetor foram avaliadas em redes artificiais e do mundo real. Verificamos que as probabilidades correlacionadas ao PageRank, o grau e a centralidade de intermediação melhoram a propagação do rumor nas redes sem escala, em comparação com o caso homogêneo. Também, modelamos a evolução do sistema considerando uma abordagem teórica de tempo discreto para a propagação heterogênea. Observou-se uma boa concordância entre os resultados teóricos e as simulações de Monte Carlo. Adicionalmente, mostramos que o espalhamento de rumores pode ser melhorado ou controlado através do ajuste das probabilidades de propagação de alguns vértices centrais.

O modelo proposto lida com a propagação de rumores considerando os contatos diretos entre os vértices. O sistema dinâmico também considerou equações em tempo discreto para a probabilidade de cada vértice ser infectado. Nossas análises generalizam, a partir de um processo de contato individual com probabilidade de propagação própria, para o caso clássico em que todos os indivíduos têm a mesma probabilidade de propagação e para qualquer topologia complexa. A solução numérica das equações analíticas superaram o custo computacional e concordam com os resultados das simulações.

Mostramos que a fração final de indivíduos informados muda ao lidar com probabilidades de transmissão heterogêneas, segundo algumas medidas de centralidade e topologias de rede. Especificamente em redes sem escala, a prevalência da informação é melhorada ou controlada quando as probabilidades de propagação são proporcionais ao PageRank, grau ou centralidades de intermediação. Portanto, as mudanças no comportamento da transmissão exercem um profundo impacto na dinâmica de propagação de acordo com a topologia da rede. Além disso, a seleção dos vértices exerce uma influência maior sobre a dinâmica do que atribuir aleatoriamente probabilidades altas de propagação. A presença de propagadores influentes no modelo heterogêneo foi correlacionada a algumas medidas de centralidade.

Nossos resultados focam na transmissibilidade individual e as interações de contato, com alguns dos comportamentos heterogêneos observados em processos reais. São fornecidas ferramentas e abordagens para uma melhor compreensão de como as ideias ou informações se espalham através das redes sociais e as condições que permitem uma ampla disseminação.

SELEÇÃO DOS VÉRTICES MAIS INFLUENTES NAS COMUNIDADES

Rumores ou informações podem se espalhar rapidamente atingindo muitos usuários em redes sociais. Modelos que auxiliem na compreensão, prevenção ou propagação de informações são de grande interesse para o setor público, privado e pesquisadores. Neste capítulo propomos uma abordagem para incrementar a difusão de informação, selecionando um conjunto de vértices mais influentes dentre as comunidades da rede. Também, analisamos a seleção dos vértices mais centrais da rede e consideramos redes artificiais e sociais, como *email*, *hamsterster*, *advogato* e *astrophysics*. Os resultados experimentais confirmam o incremento na fração final de indivíduos informados ao aplicar uma abordagem baseada em comunidades, ou seja, selecionar os vértices mais centrais das comunidades incrementa muito mais o alcance da propagação do que selecionar os vértices mais centrais da rede toda. O texto a seguir baseia-se nos resultados publicados em (VEGA-OLIVEROS; BERTON, 2015).

O capítulo está estruturado como segue: na Seção 5.1 apresentamos a motivação e lacuna na área relacionada com seleção de um conjunto de vértices influentes. A Seção 5.2 traz as abordagens e trabalhos relacionadas com o tema pesquisado. Na Seção 5.3 apresentamos a proposta de seleção dos vértices mais influentes da estrutura de comunidades da rede. As redes complexas consideradas e a metodologia empregada neste trabalho são apresentadas nas Seção 5.4 e Seção 5.5 respectivamente. Os resultados são apresentados na Seção 5.6. Finalmente, as considerações finais do capítulo estão na Seção 5.7.

5.1 Motivação

A modelagem de processos de propagação ou difusão em redes sociais tem recebido muita atenção recentemente, pois permite compreender como uma doença pode ser controlada ou como uma informação se difunde entre os indivíduos de uma rede. Modelos que descrevem

a evolução dos rumores em redes podem ser adaptados para analisar a propagação de *spam* na Internet, publicidade e *marketing*, ideologias políticas ou notícias de avanços tecnológicos, entre outros (CASTELLANO; FORTUNATO; LORETO, 2009). Em tais casos, a representação em redes complexas permite a análise de modelos tradicionais e da estrutura heterogênea, que tem forte influência nos processos de difusão (MORENO; NEKOVEE; PACHECO, 2004; BARRAT; BARTHÉLEMY; VESPIGNANI, 2008; CASTELLANO; FORTUNATO; LORETO, 2009). Alguns indivíduos podem ter uma influência maior do que outros segundo a estrutura da rede. A maioria dos trabalhos na área tem se concentrado na identificação dos vértices que sejam mais influentes (KITSACK *et al.*, 2010; LAWYER, 2012; PEI; MAKSE, 2013; HÉBERT-DUFRESNE *et al.*, 2013). Acredita-se que os vértices mais centrais podem convencer o maior número de indivíduos na rede. No entanto, a seleção de mais de um vértice central não necessariamente incrementa a fração esperada de indivíduos informados. Pode existir interferências ou sobreposições na influência, que levaria a resultados piores do que a seleção uniformemente aleatória de vértices.

Por outro lado, nas redes reais as conexões entre os vértices podem estar associadas a uma estrutura modular. Essas estruturas simbolizam grupos ou comunidades de vértices que estão mais densamente conectadas entre si. Assim, a identificação dessas comunidades pode ser útil para identificar características dos indivíduos que pertencem à comunidade. Isto é devido a uma maior semelhança entre os vértices de uma mesma comunidade, os quais tendem a ser parecidos segundo o fenômeno de homofilia (WENG; MENCZER; AHN, 2013; LOPES; PINTO; FRANCISCO, 2016). Esse conceito indica que as pessoas estão mais predispostos a adotar e disseminar informações que provém da própria comunidade, por um fator psicológico/social de afinidade com aquele grupo de indivíduos (STIEGLITZ; DANG-XUAN, 2013; WENG; MENCZER; AHN, 2013; LOPES; PINTO; FRANCISCO, 2016).

Portanto, é importante considerar as características topológicas da rede para a correta identificação dos vértices que melhor propaguem uma informação. Por exemplo, consideremos uma empresa que deseja comercializar um novo produto na blogosfera. A empresa poderia selecionar três indivíduos muito influentes desta rede social (blogueiros com milhares de acesso) para anunciar seu produto. No entanto, esses influenciadores podem ser populares dentro do mesmo grupo de pessoas. Por outro lado, se a estratégia é identificar as três principais comunidades da rede e selecionar os indivíduos mais influentes de cada comunidade, a empresa pode atingir grupos variados de usuários e assim aumentar o alcance da campanha publicitária.

5.2 Trabalhos relacionados

Caracterizar os indivíduos que são mais influentes no processo de propagação tem atraído a atenção dos pesquisadores (RICHARDSON; DOMINGOS, 2002; KEMPE; KLEINBERG; TARDOS, 2003; KITSACK *et al.*, 2010; PEI; MAKSE, 2013). A identificação dos melhores

propagadores da rede apresenta dois problemas de pesquisa (PEI; MAKSE, 2013): *i*) caracterizar de modo geral os indivíduos que são mais influentes no processo de propagação (KITSACK *et al.*, 2010) e *ii*) encontrar o conjunto de vértices que maximizam a influência ou capacidade de propagação da rede (RICHARDSON; DOMINGOS, 2002).

No primeiro caso, a abordagem convencional para descrever os vértices mais influentes é realizando-se uma análise microscópica na rede. Os vértices são classificados considerando suas propriedades topológicas, para depois serem ordenados ou correlacionados de modo a generalizar sua capacidade de propagação (KITSACK *et al.*, 2010; HÉBERT-DUFRESNE *et al.*, 2013; PEI; MAKSE, 2013). Por exemplo, alguns trabalhos afirmam que a seleção das sementes ou propagadores iniciais de acordo com alguma característica topológica, como o grau, centralidade de intermediação, ou PageRank, promove um incremento na influência (KITSACK *et al.*, 2010; HUANG *et al.*, 2014; KANDHWAY; KURI, 2016; VEGA-OLIVEROS; COSTA; RODRIGUES, 2017b). Isto significa que não existe um consenso geral sobre qual é a medida que melhor descreva os vértices mais influentes.

Alguns pesquisadores afirmam que os *hubs* são mais representativos para influenciar outros vértices (PASTOR-SATORRAS; VESPIGNANI, 2001; ALBERT; BARABÁSI, 2002). Os vértices com maior grau são mais eficientes para espalhar a informação porque, no geral, os *hubs* não tendem a estar conectados e assim podem alcançar um maior número de vértices (KITSACK *et al.*, 2010). Outros trabalhos mostraram que os vértices mais influentes são aqueles com a maior centralidade de intermediação, pois intermediam a comunicação entre grupos de vértices, o que aumenta sua importância (HÉBERT-DUFRESNE *et al.*, 2013). Segundo os autores, a centralidade de intermediação também seria um bom descritor de propagadores influente nas comunidades. O PageRank também é considerado uma medida muito precisa para caracterizar os vértices mais influentes (CATALDI; Di Caro; SCHIFANELLA, 2010; VEGA-OLIVEROS; COSTA; RODRIGUES, 2017b). A razão é que emprega o conceito de caminhada aleatória sobre a rede e vértices com valores mais elevados possuem maior probabilidade de ser visitados.

Pode-se pensar que o problema de encontrar um conjunto de propagadores iniciais é resolvido selecionando os vértices que são mais centrais segundo o primeiro caso. No entanto, a seleção dos vértices iniciais segundo o nível global de centralidade pode produzir uma sobreposição na influência entre essas sementes (KITSACK *et al.*, 2010; PEI; MAKSE, 2013; KEMPE; KLEINBERG; TARDOS, 2015). De fato, foi mostrado que o problema de encontrar um grupo de vértices que maximize a propagação é um problema de otimização NP-difícil (KEMPE; KLEINBERG; TARDOS, 2003).

Desse modo, Kempe, Kleinberg e Tardos (2003) propõem o método baseado na escolha pelo método guloso (chamamos aqui de Greedy) para obter η propagadores iniciais que maximizam a influência da difusão. Os autores adotaram uma abordagem de otimização discreta e provaram que o problema de otimização é NP-difícil. Na proposta foi considerado o modelo de propagação de cascata independente e ponderada (IC) que possui apenas dois estados, diferentes

do modelo SIR. O método avalia um vértice por vez para ser adicionado no conjunto de sementes selecionadas. O novo vértice será aceito se for aquele que mais incrementa a média de difusão. Porém, esta abordagem tem o problema de ter um custo computacional muito alto, embora novas pesquisas tentem otimizar o seu desempenho (CHEN; WANG; YANG, 2009; KEMPE; KLEINBERG; TARDOS, 2015).

Contudo, trabalhos recentes verificaram que para melhorar a eficiência da propagação, uma boa estratégia é distribuir as sementes entre as comunidades (GALSTYAN; MUSOYAN; COHEN, 2009; WANG *et al.*, 2010; WENG; MENCZER; AHN, 2013). Galstyan, Musoyan e Cohen (2009) afirmam que se a estrutura de comunidades não for considerada, então somente soluções subótimas podem ser obtidas. Isso acontece porque os vértices pertencentes à mesma comunidade são susceptíveis de serem mais semelhantes entre si e compartilhar o mesmo conjunto de vizinhos. Os autores empregaram uma abordagem gulosa para selecionar as sementes da menor comunidade e verificaram que é possível causar uma ativação em cascata global, mesmo para um pequeno número de sementes. No entanto, os resultados são restritos a redes aleatórias de duas comunidades. Wang *et al.* (2010) propuseram um algoritmo guloso baseado em comunidades para definir os η vértices mais influentes. A idéia foi encontrar a divisão de comunidades da rede e, por meio de um algoritmo de programação dinâmica, selecionar gradualmente a comunidade a partir da qual o próximo nó influente será selecionado. O método possui um custo computacional relativamente alto, embora seja mais rápido que o algoritmo Greedy para encontrar nós influentes. Em uma abordagem semelhante, Cao *et al.* (2011) transformaram o problema de maximização da influência em um problema de alocação ótima de recursos nas comunidades da rede. Eles assumiram que as comunidades estão desconectadas. Em seguida, o método seleciona η candidatos de cada comunidade segundo a centralidade de grau. No final, um algoritmo de programação dinâmica identifica quais serão os vértices a ser considerados como sementes.

5.3 Seleção de propagadores influentes segundo as comunidades

Uma comunidade é um grupo de vértices que tem mais conexões entre si do que com vértices de outros grupos (ver Seção 2.1.4.3). Nas redes sociais, as comunidades representam pessoas que compartilham afinidades, o que reforça os laços e o fenômeno de homofilia (WENG; MENCZER; AHN, 2013; LOPES; PINTO; FRANCISCO, 2016). Esta condição é a razão pela qual informações ou sentimentos se propagam melhor dentro das comunidades, tornando as pessoas mais abertas à informação compartilhada por seus pares (WENG; MENCZER; AHN, 2013; LOPES; PINTO; FRANCISCO, 2016).

Assim, propomos uma nova abordagem para a seleção dos propagadores iniciais a partir dos vértices mais centrais dentro de cada comunidade. Foi verificado que as medidas

Tabela 3 – Propriedades estruturais e resultados da detecção de comunidades nas redes adotadas. O melhor valor de modularidade Q e o número de comunidades dessa divisão, segundo o algoritmo *fastgreedy*, estão na última coluna

Rede	N	$\langle k \rangle$	$\langle \ell \rangle$	$\tilde{H}$	ρ	fastgreedy	
						Q	Nc
<i>email</i>	1133	9.62	3.60	0.45	0.01	0.49	16
<i>hamsterster</i>	2000	16.1	3.58	0.48	0.02	0.46	57
<i>advogado</i>	5054	15.6	3.27	0.40	-0.09	0.34	49
<i>astrophysics</i>	14845	16.1	4.79	0.38	0.23	0.63	1172

de centralidade nas comunidades (escala mesoscópica) se comparam favoravelmente à escala global (HÉBERT-DUFRESNE *et al.*, 2013). Muitos métodos podem detectar as comunidades na rede (WANG *et al.*, 2010), mas aqui empregaremos o algoritmo *fastgreedy* que é um método rápido e preciso (NEWMAN, 2004; FORTUNATO; HRIC, 2016). As principais η comunidades da rede são detectadas, onde cada uma é definida pelo subgrafo $G_c(V_c, E_c)$ e $\sum_{c=1}^{\eta} |G_c(V_c, E_c)| = |V|$. A seguir, para cada comunidade calculamos uma medida de centralidade e selecionamos o seu vértice mais central. Portanto, obtemos η vértices que são o conjunto de sementes iniciais. Dado que os vértices com maior centralidade são considerados os mais adequados para a difusão na rede, selecionamos então o vértice mais central ou importante de cada comunidade. Portanto, esses vértices tem maior influência em sua própria comunidade e a sobreposição de influências na população é minimizada.

5.4 Redes consideradas

Neste trabalho consideramos as bases *email* (GUIMERA *et al.*, 2003), *advogado* (KUNEGIS, 2017; MASSA; SALVETTI; TOMASONI, 2009), *astrophysics* (NEWMAN, 2001) e *hamsterster* (KUNEGIS, 2017). Todas elas foram assumidas como redes não direcionadas sem peso e também foi considerado o maior componente para as simulações. As propriedades estruturais das redes estão resumidas na Tabela 3, com o respectivo número de vértices N , o grau médio $\langle k \rangle$, a média do comprimento dos caminhos mais curtos $\langle \ell \rangle$, entropia normalizada $\tilde{H}$ e correlação de Pearson ρ . Além disso, o melhor valor de modularidade Q e número de comunidades segundo essa divisão NC são calculados segundo o algoritmo de *fastgreedy*.

5.5 Metodologia

Conforme definido na Seção 3.2, pela natureza desse problema modificamos a configuração inicial considerando $\psi(0) = 1 - \eta/N$, $\phi(0) = \eta/N$ e $\varphi(0) = 0$, onde η representa as sementes ou os propagadores iniciais. Assim, cada simulação começa com a seleção do conjunto de η vértices.

As sementes serão selecionadas a partir da divisão de comunidades da rede, sendo os propagadores iniciais os vértices mais centrais de cada comunidade. A divisão em comunidades pode ser calculada por algum método divisivo ou aglomerativo (Seção 2.1.4.3) e aqui utilizamos o algoritmo *fastgreedy*. O método proposto é detalhado como segue: Primeiro, dado um número requerido de η propagadores iniciais, detectamos as η comunidades principais da rede pelo algoritmo *fastgreedy*. Depois, cada comunidade é isolada, o que produz η componentes. O processo de isolamento consiste em manter as arestas intra-comunidades e remover as conexões entre as comunidades. Para cada comunidade isolada, uma medida de centralidade específica é calculada para todos os vértices. Ao final, as η sementes são selecionadas onde cada uma é a mais central da sua comunidade. Retomamos a rede completa original, as η sementes e os parâmetros para executar as simulações correspondentes.

Nas simulações consideramos três opções de escolha das sementes: (i) selecionando η indivíduos aleatoriamente como sementes iniciais da rede; (ii) selecionando os η indivíduos com os valores mais altos de uma dada medida de centralidade, ignorando a estrutura de comunidades; e finalmente, (iii) pelo método que propomos baseado em comunidades. As medidas de centralidade selecionadas foram o grau (DG), centralidade de intermediação (BE) e PageRank (PR).

5.6 Resultados

Nesta seção, são apresentados os resultados da propagação de rumores em uma rede artificial sem escala e quatro redes sociais. Avaliamos o impacto que o comportamento dos propagadores têm na dinâmica de difusão na rede. Em seguida, analisamos os resultados sobre a seleção dos propagadores iniciais segundo o método proposto por comunidades, os vértices mais centrais da rede e de forma aleatória com probabilidade uniforme.

5.6.1 Impacto do comportamento dos propagadores

A fim de exemplificar as diferenças entre as abordagens, analisamos a fração final de contidos segundo as taxas β e μ , ambas com valores no intervalo $(0, 1]$ (Figura 18). Consideramos o modelo *MT* em uma rede artificial sem escala de tamanho $N = 1000$ e $\langle k \rangle = 8$. Para cada tupla de valores (β, μ) foram feitas 100 realizações, escolhendo aleatoriamente 1% dos vértices como propagadores iniciais, e para cada escolha foram feitas 50 iterações.

Observamos que o *CP*, é o processo mais restritivo com a dinâmica de difusão, executando por definição um menor número de tentativas de contágio. Desse modo, a fração final de contidos é menor em comparação aos outros processos. Assim, este tipo de contato assemelha-se mais a uma propagação do tipo “boca-a-boca”.

Ao contrário, o *RP*, atinge mais de 80% de indivíduos informados na região de valores de $\beta \geq 0.3$, sem importar o valor que μ possa ter. Desse modo, trata-se de um processo

Figura 18 – Modelo [Maki e Thompson \(1973\)](#) para propagação de rumores em uma rede artificial BA de $N = 1000$ vértices, $\langle k \rangle = 8$ e $\eta = 1\%|V|$ de sementes iniciais. A barra de cor simboliza a fração final de contidos: (a) processo reativo (RP); (b) processo truncado (TP); e (c) processo de contato (CP)

Fonte: Elaborada pelo autor.

Figura 19 – Modelo MT-TP em uma rede sem escala artificial com $N = 1000$ e $\langle k \rangle = 8$. A fração final de indivíduos informados é definida nas barras de cor. A seleção de $\eta = 4\%|V|$ de sementes foi conforme: (a) o método proposto baseado em comunidades considerando os vértices com melhor centralidade de PageRank; (b) seleção aleatória uniforme de η vértices; e (c) os η nós com a maior centralidade PR da rede. As linhas brancas contínuas à esquerda representam combinações de parâmetros (β, μ) que atingiram 35% de indivíduos contidos. As linhas brancas tracejadas mostram as combinações que alcançaram 60% de indivíduos informados.

Fonte: Adaptada de [Vega-Oliveros e Berton \(2015\)](#).

menos restritivo, privilegiando uma difusão viral e quase completa na rede. Por esse motivo o comportamento *RP* é mais indicado para simular propagações do tipo *broadcasting*.

O *TP*, está mais relacionado com o cenário de contato em redes, onde a localização e características topológicas dos vértices têm maior impacto na difusão. Este processo apresenta gráficos semelhantes ao processo *CP*, com resultados perto de 60% quando $\beta \approx \mu$, mas o contato não é tão restrito quanto ao *CP*. Por esta razão, adotamos a abordagem *TP* como o comportamento de difusão dos propagadores nas análises posteriores.

5.6.2 Análises da seleção de múltiplas sementes

A fração final de vértices informados ($\varphi(\infty)^{\mathbf{S}}$) foi calculada sobre 100 execuções para cada combinação de sementes iniciais ($\mathbf{S}$) e parâmetros. Esta média representa a capacidade de propagação alcançada pelas sementes selecionadas. Foi avaliada a relação entre os parâmetros e a seleção dos propagadores iniciais em uma rede sem escala artificial. Nesta primeira fase, apenas a medida de PageRank foi considerada para encontrar as sementes nas comunidades e em toda a rede. Uma fração de $\eta = 4\%|V|$ de propagadores iniciais foi definida para uma rede sem escala *BA* (BARABÁSI; ALBERT, 1999) de tamanho $N = 1000$, $\langle k \rangle = 8$, $\langle g \rangle = 3.19$, $\tilde{H} = 0.33$ e $\rho = -0.04$.

A capacidade de propagação $\varphi(\infty)^{\mathbf{S}}$ foi afetada de acordo com a seleção das sementes (Figura 19). As linhas sólidas e tracejadas representam a combinação de parâmetros β e μ que obtiveram 35% e 60% de vértices contidos, respectivamente. Observamos que essas curvas apresentam um padrão linear bem definido, o que significa que qualquer proporção de $\lambda = \beta/\mu$ obtém resultados equivalentes de $\varphi(\infty)^{\mathbf{S}}$. Observamos que a seleção das sementes por comunidades melhorou a capacidade de difusão na rede em comparação com as sementes aleatórias (Figure 19(b)) ou as sementes que são os vértices mais centrais da rede (Figure 19(c)). Este resultado é corroborado pelo aumento da inclinação das linhas brancas. No entanto, é evidenciado uma pequena queda na inclinação das linhas de *BST* (caso vértices mais centrais) em comparação com *RANDOM* (caso uniformemente aleatório).

Analisamos o impacto do tamanho de $|\mathbf{S}| = \eta$ e as medidas de centralidade no processo de propagação. Variamos o número de comunidades e sementes de 2 a 50 e fixamos $\beta = 0.3$ e $\mu = 0.2$ para todas as simulações. As redes sociais e o modelo de propagação MT-TP foram considerados na análise (Figura 20). A seleção aleatória de propagadores iniciais (pontos azuis, *RANDOM*) ou vértices mais centrais da rede (pontos pretos, *BST*) das centralidades *DG*, *BE* ou *PR*, produzem valores de capacidade de propagação ($\varphi(\infty)^{\mathbf{S}}$) que variam pouco. Em alguns casos, a seleção aleatória de sementes atingiu valores de $\varphi(\infty)^{\mathbf{S}}$ maiores do que a seleção dos melhores vértices da rede (*BST*). Observa-se que para os métodos *BST* existe uma tendência de quanto maior o número de propagadores iniciais, menores os valores de $\varphi(\infty)^{\mathbf{S}}$, resultado não esperado ou contra-intuitivo.

Por outro lado, quando a detecção da comunidade foi realizada e os vértices com maiores valores de centralidade *DG*, *BE* ou *PR* de cada comunidade (pontos vermelhos, *COM*), foram selecionados, a capacidade de propagação obteve os maiores resultados. Portanto, mais vértices foram informados na rede mediante a seleção por comunidade. Contudo, chama a atenção o caso da rede *advogado* (Figura 26d), a qual apresentou um pequeno pico nos resultados para os métodos *BST*. Este fenômeno é abordado e estudado no Capítulo 6.

Em termos das medidas topológicas, observamos que os vértices com maior centralidade de PageRank nas comunidades (*COM-PR*) obtiveram em média melhores resultados de propa-

Figura 20 – Capacidade de propagação do modelo MT-TP nas quatro redes reais consideradas, dada a seleção de sementes por comunidades (pontos vermelhos), melhor ranqueados ou centrais (pontos pretos) e aleatoriamente (pontos azuis).

Fonte: Adaptada de [Vega-Oliveros e Berton \(2015\)](#).

Tabela 4 – Média das capacidades de propagação para os valores de $\eta \in [1, 50]$, para cada rede social e método de seleção: (segunda coluna maior) selecionando os vértices mais centrais das comunidades; (Terceira coluna maior), selecionando os vértices mais centrais da rede; e (última coluna maior) selecionando-os aleatoriamente. Em negrito destacamos os melhores resultados entre as métricas. As medidas adotadas foram a centralidade de intermediação (*BE*), grau (*DG*) e PageRank (*PR*).

Rede	Comunidades			Mais centrais			Seleção aleatória
	<i>BE</i>	<i>DG</i>	<i>PR</i>	<i>BE</i>	<i>DG</i>	<i>PR</i>	
<i>email</i>	0.6065	0.6105	0.6150	0.5880	0.5840	0.5884	0.6023
<i>hamsterster</i>	0.5485	0.5693	0.5728	0.5306	0.5226	0.5271	0.5273
<i>advogato</i>	0.4077	0.4102	0.4112	0.3993	0.3958	0.4007	0.3805
<i>astrophysics</i>	0.5417	0.5398	0.5415	0.5321	0.5278	0.5301	0.5337

gação (Tabela 4). Mesmo na seleção dos vértices melhor ranqueados, o PageRank foi notável. Outro ponto importante é que muitas vezes, a seleção aleatória uniforme do propagadores iniciais foi melhor do que a seleção dos vértices mais centrais (melhor classificados) da rede. Isso é contrário ao que é atualmente esperado e adotado em campanhas de marketing, por exemplo. Os vértices mais centrais da rede produzem em conjunto uma influência baixa, obtendo valores de

fração final de vértices informados menores do que se selecionarmos vértices aleatoriamente. No entanto, a seleção de propagadores iniciais por comunidades mostrou, independentemente da medida de centralidade, bons resultados.

5.7 Considerações finais

Neste capítulo foi apresentado um método para a seleção de vértices influentes baseado na estrutura de comunidades da rede. Primeiro, analisamos o impacto do comportamento dos propagadores na dinâmica de difusão e observamos que o Processo Truncado (TP) e o Processo de Contato (CP) são mais adequados para simular a propagação de informação em redes sociais. Aplicamos a detecção de comunidades e definimos como vértices mais influentes aqueles com maior centralidade nas comunidades. Os resultados experimentais em uma rede sem escala e quatro redes sociais confirmaram o aumento na fração final de indivíduos informados com a abordagem proposta. Além disso, verificou-se que a centralidade do PageRank nas comunidades é a melhor escolha em termos de eficiência e fração final de vértices informados.

Apresentamos nossa proposta para selecionar os propagadores iniciais segundo a divisão de comunidades do algoritmo *fastgreedy*. Ainda existe o problema em aberto relacionado com a definição do que considera-se uma comunidade e qual seria a divisão ideal. No entanto, variamos o número de comunidades de 2 a 50 e, em geral, (para cada divisão de comunidades) a nossa proposta obteve melhores resultados versus a propagação sem considerar a estrutura da comunidade ou considerando os vértices mais centrais da rede.

IMPACTO DA ASSORTATIVIDADE NA MAXIMIZAÇÃO DE INFORMAÇÃO

Neste capítulo apresentamos uma análise do papel da correlação de grau no problema da maximização da influência. Realizamos exaustivas simulações em oito redes reais e seis redes artificiais sem escala. Variamos o nível de assortatividade em redes artificiais e o número de sementes iniciais, de 2 a no mínimo 10% do tamanho da rede. Buscando considerar a topologia da rede e melhorar o custo computacional, adotamos a abordagem simples de selecionar as sementes das comunidades da rede para maximizar a difusão da informação. Nosso método seleciona como sementes iniciais os vértices mais centrais dentro de cada comunidade. Também investigamos o desempenho de dois métodos amplamente conhecidos: (i) a seleção dos vértices mais centrais e (ii) o algoritmo de otimização guloso. Para cada método, realizamos uma análise da velocidade da difusão, do pico máximo de propagadores e a distribuição das sementes de acordo com a estrutura de comunidades.

Além disso, examinamos os resultados da maximização da influência com o teste estatístico Wilcoxon e o pós-hoc Nemenyi-Freidman. Como resultado, verificamos que a assortatividade desempenha um papel significativo na dinâmica do problema de maximização da influência. Por exemplo, em redes disassortativas quando o número de sementes é menor que 1% do tamanho de rede, selecionar os vértices mais centrais como espalhadores iniciais produz um pico de influência que é maior do que os resultados do método guloso. Em redes assortativas, a seleção dos vértices mais centrais por comunidades aumenta a difusão da informação mais intensamente do que a seleção dos indivíduos mais centrais de toda a rede. Adicionalmente, os resultados de maximização do método proposto não têm diferenças estatísticas com a abordagem gulosa, mas a propagação da informação é mais rápida. Consequentemente, métodos que não levam em conta as propriedades estruturais da rede, como no caso da otimização gulosa, podem levar a resultados sub-ótimos.

O capítulo está estruturado como segue: na Seção 6.1 apresentamos a motivação e lacunas

na área relacionada com o impacto da correlação grau-grau tem nas dinâmicas de propagação em redes. A Seção 6.2 traz a definição e trabalhos relacionadas com o problema de maximização da propagação de informação ou influência. As redes complexas consideradas e a metodologia empregada neste trabalho são apresentadas nas Seção 6.3 e Seção 6.4 respectivamente. Os resultados do impacto da assortatividade nas redes artificiais e reais e a distribuição das sementes são apresentadas na Seção 6.5. Na Seção 6.6 apresentamos uma análise estatística dos resultados de propagação dos métodos em relação ao tamanho do conjunto de sementes iniciais, em até 10% os vértices da rede. Finalmente, as considerações finais do capítulo estão na Seção 6.7.

6.1 Motivação

A maximização da difusão de informação, ou maximização da influência, é de fundamental importância em cenários como propaganda viral, campanhas políticas entre outras (KEMPE; KLEINBERG; TARDOS, 2003; MORONE; MAKSE, 2015). Muitas abordagens têm sido propostas para o IMP, (KEMPE; KLEINBERG; TARDOS, 2003; CASTELLANO; FORTUNATO; LORETO, 2009; MORONE; MAKSE, 2015). Por exemplo, alguns trabalhos afirmam que a seleção dos propagadores iniciais (sementes) segundo característica topológicas, como o grau, a centralidade de intermediação, ou PageRank, promovem uma maximização da influência (KITSAK *et al.*, 2010; HUANG *et al.*, 2014; KANDHWAY; KURI, 2016; VEGA-OLIVEROS; COSTA; RODRIGUES, 2017b). No entanto, a seleção do conjunto de sementes dos vértices mais centrais não implica necessariamente na maximização da propagação, pois pode ocorrer uma sobreposição de influência entre as sementes (PEI; MAKSE, 2013; KEMPE; KLEINBERG; TARDOS, 2015). Um dos métodos mais destacado para resolver o problema de maximização de influência é uma abordagem gulosa *Hill climbing* de otimização, que assegura uma resposta próxima de 63% do ótimo para várias classes de modelos de influência. Morone e Makse (2015) mapearam o problema na percolação ideal em redes aleatórias para identificar o conjunto mínimo de propagadores influentes. Os autores verificaram que o conjunto mínimo de propagadores influentes são dadas pelos nós cuja remoção fragmenta a rede em muitos componentes desconexos. Embora esses trabalhos tenham avançado no estudo da maximização da influência, eles desconSIDERAM padrões de conexões, como a correlação grau-grau e a estrutura de comunidades, que desempenham um papel importante no processo de propagação (VEGA-OLIVEROS; COSTA; RODRIGUES, 2017a). Os padrões das conexões têm um impacto fundamental na dinâmica de propagação (PASTOR-SATORRAS *et al.*, 2015). Entre os vários padrões não triviais que podem ser encontrados em redes complexas, a assortatividade desempenha um papel destacado na dinâmica de propagação (ná; PASTOR-SATORRAS; VESPIGNANI, 2003). A assortatividade, ou correlação de grau, é uma propriedade da rede na qual vértices com características semelhantes, como grau, tendem a estar conectados. Trabalhos anteriores verificaram que as epidemias se difundem mais rapidamente em redes assortativas, mas o alcance é maior em estruturas disassortativas (KISS; GREEN; KAO, 2008). A assortatividade também afeta o limiar de propagação (ná;

PASTOR-SATORRAS; VESPIGNANI, 2003) e o tempo de difusão (BERTOTTI; BRUNNER; MODANESE, 2016).

O papel da estrutura de comunidades na propagação de memes em redes sociais foi analisado em (WENG; MENCZER; AHN, 2013). Eles apresentam um método para prever a viralidade dos memes com base em padrões iniciais de contágio nas comunidades. No entanto, não foi abordado o problema de maximização da influência. Zhang *et al.* (2013) identificaram os η vértices mais influentes aplicando o algoritmo de agrupamento k-medoid em uma matriz de similaridade obtida a partir de um processo de percolação. Embora tenham fornecido evidências empíricas para o modelo IC para pequenos conjuntos de dados, o método tem alto custo computacional. Recentemente, Hosseini-Pozveh, Zamanifar e Naghsh-Nilchi (2016) introduziu uma medida de centralidade de proximidade baseada na influência para encontrar η vértices candidatos por comunidade, onde η é o tamanho do conjunto de sementes iniciais esperado. Posteriormente, o conjunto de sementes finais é obtido através de calcular algumas heurísticas de modo a filtrar os melhores candidatos.

Observamos que nenhum desses trabalhos aborda o impacto da assortatividade no problema de maximização da difusão de informação. Além disso, esses métodos possuem um custo computacional alto e consideram um número relativamente baixo de sementes, até $\eta = 50$ vértices. Os modelos de rumores clássicos também não são abordados por esses métodos, embora os modelos de Daley e Kendall (1964) e Maki e Thompson (1973) sejam frequentemente usados para modelar a dinâmica de rumores em redes (ZANETTE, 2001b; MORENO; NEKOVEE; VESPIGNANI, 2004; NEKOVEE *et al.*, 2007; BORGE-HOLTHOEFER; MORENO, 2012). No caso do problema de maximização da influência, a assortatividade ainda não foi considerada (ver por exemplo (KEMPE; KLEINBERG; TARDOS, 2015; PASTOR-SATORRAS *et al.*, 2015)). Assim, neste capítulo estudamos o impacto da assortatividade sobre a maximização da influência e a velocidade de propagação, considerando a seleção de várias sementes (pelo menos 10% do tamanho da rede).

6.2 O problema de maximização de informação

O problema de maximizar a influência ou a difusão de informação visa encontrar o conjunto de η vértices que maximizem a propagação de uma informação. A informação propagada a partir destes vértices abrange o número máximo de nós na rede. Mais especificamente, este problema é encarado como a seleção de um conjunto de propagadores influentes, ou sementes iniciais, que acionam a maior cascata de novos seguidores de acordo com algum modelo de propagação (RICHARDSON; DOMINGOS, 2002; KEMPE; KLEINBERG; TARDOS, 2003). O problema de encontrar o melhor conjunto de sementes iniciais é NP-hard na maioria dos modelos de propagação (KEMPE; KLEINBERG; TARDOS, 2015), o que o torna um desafio para pesquisadores nas áreas de física, computação e ciências sociais. Assim, apesar do uso de

estratégias heurísticas para identificar os propagadores mais influentes, o problema permanece em aberto.

Como exemplo motivacional, consideremos uma empresa que possui dados sobre uma importante rede social e queira comercializar um novo produto. De acordo com os dados, é possível estimar a influência que alguns indivíduos têm sobre outros. O objetivo é selecionar um grupo de indivíduos iniciais que desencadeiem a maior cascata de influência, na qual as pessoas recomendarão o produto e aumentará o número de seguidores ou consumidores. Para selecionar esses influenciadores iniciais –a companhia pode motivá-los dando amostras grátis do produto ou pagando pela participação–, a empresa poderia selecionar segundo um critério global de importância de acordo com os dados da rede social, por exemplo, os usuários com o maior número de conexões. No entanto, esses influenciadores podem interagir no mesmo clique ou grupo de indivíduos, o que pode produzir uma situação de sobreposição de influência. Por outro lado, se a estratégia é identificar as principais comunidades da rede e selecionar os indivíduos influentes de cada comunidade, a empresa conseguiria um grupo variado de usuários. Finalmente, a empresa pode empregar uma estratégia de força bruta, onde tenta descobrir quase todos os conjuntos possíveis de influenciadores iniciais possíveis que produza a melhor configuração. Embora se espera que esta opção gulosa obtenha bons resultados, tem um custo muito alto e os resultados podem chegar a não ser significativamente diferentes das abordagens topológicas que consideram a estrutura da rede.

Assim, o problema de maximização da influência (IMP) procura o conjunto de vértices S que contêm $|S| = \eta$ sementes que maximizam um processo de difusão ou influência na rede. O impacto de um dado conjunto de vértices ($\sigma(S)$) corresponde à fração esperada de vértices que receberam a informação no final do processo.

6.3 Redes consideradas

Realizamos extensas simulações numéricas em várias redes artificiais e do mundo real, avaliando o impacto da correlação de grau no problema de maximização de influência. Selecionamos as redes do mundo real de acordo com o nível de correlação grau-grau, onde metade delas são disassortativas e as outras assortativas. As redes disassortativas adotadas são a *Google+*, *internet*, *caida* e *advogato*. As redes assortativas adotadas são *email*, *hamsterster*, *PGP* e *astrophysics*. Assumimos que todas as redes são não direcionadas e não ponderadas e apenas o maior componente de rede foi considerado. A descrição das redes reais está na Seção 3.3.

No caso das redes artificiais utilizamos a abordagem proposta por [Xulvi-Brunet e Sokolov \(2005\)](#) para controlar a correlação grau-grau em redes [Barabási e Albert \(1999\)](#) (BA). Adotamos a rede BA porque apresentam distribuição de grau sem escala, como as redes sociais, e correlação nula. Então, ao religar as arestas conseguimos redes conexas com diferentes níveis de assortatividade ([Xulvi-Brunet; Sokolov, 2005](#)). Observe que nas redes assortativas, vértices

Tabela 5 – Propriedades estruturais das redes consideradas. ρ é o coeficiente de assortatividade, $|V|$ o tamanho da rede, $\langle k \rangle$ o grau médio, $\langle \ell \rangle$ é o comprimento médio dos caminhos mínimos e $\langle C_i \rangle$ e a média do coeficiente de agrupamento. Os parâmetros relacionados à comunidade são a modularidade Q e o número de comunidades N_c .

Rede	ρ	$ V $	$\langle k \rangle$	$\langle \ell \rangle$	$\langle C_i \rangle$	fastgreedy Q	N_c
<i>BA</i>	-0.43	1000	11.9	2.96	0.017	0.26	8
<i>BA</i>	-0.31	1000	11.9	2.87	0.028	0.25	9
<i>BA</i>	-0.21	1000	11.9	2.86	0.031	0.25	9
<i>BA</i>	0.02	1000	11.9	2.91	0.035	0.25	10
<i>BA</i>	0.11	1000	11.9	2.94	0.034	0.25	11
<i>BA</i>	0.34	1000	11.9	3.11	0.026	0.27	8
<i>Google+</i>	-0.39	23613	3.32	4.03	0.174	0.74	33
<i>Internet</i>	-0.20	22963	4.22	3.84	0.231	0.63	57
<i>Caida</i>	-0.20	26475	4.03	3.87	0.208	0.64	43
<i>Advogato</i>	-0.09	5054	15.6	3.27	0.253	0.34	49
<i>email</i>	0.01	1133	9.62	3.60	0.220	0.49	16
<i>Hamsterster</i>	0.02	2000	16.1	3.58	0.539	0.46	57
<i>PGP</i>	0.23	10680	4.55	7.48	0.266	0.85	179
<i>Astrophysics</i>	0.23	14845	16.1	4.79	0.638	0.63	1172

Figura 21 – Representação gráfica circular de uma rede BA com a correlação de graus modificada pela abordagem de [Xulvi-Brunet e Sokolov \(2005\)](#): (a) BA disassortiva com $\rho = -0.43$ e (b) BA assortativa com $\rho = 0.34$.

(a) Rede disassortativa

(b) Rede assortativa

Fonte: Adaptada de [Vega-Oliveros, Costa e Rodrigues \(2017a\)](#).

com grau similar tendem a estar conectados, e nas disassortativas, vértices com alta centralidade de grau estão conectados com nós de baixo grau. Figura 21 apresenta exemplos de redes assortativas e disassortativas.

As propriedades estruturais das redes são resumidas na Tabela 5, com a respectiva assortatividade ρ (Seção 2.1.4.1), número de vértices $|V|$, grau médio $\langle k \rangle$, comprimento médio dos caminhos mínimos $\langle \ell \rangle$ e a média do coeficiente de agrupamento $\langle C_i \rangle$. Além disso, o maior

valor de modularidade Q e número de comunidades NC foi identificado pelo algoritmo *fastgreedy*. Podemos ver que as redes sociais são mais modulares do que as redes artificiais. No entanto, a maioria das redes apresentam valores semelhantes de comprimento médio dos caminhos mínimos ($\langle \ell \rangle$).

6.4 Metodologia

Consideramos a mesma configuração geral definida na Seção 3.2 com o processo truncado (TP) como comportamento dos propagadores (BORGES-HOLTHOEFER; MORENO, 2012). Podemos calcular a fração final de indivíduos informados ($\varphi(\infty)\mathbf{S}$), isto é, $\varphi(\infty)\mathbf{S} = |R(\infty)|/|V|$ ou $\varphi(\infty)\mathbf{S} = 1 - |S(\infty)|/|V|$. No entanto, $\varphi(\infty)\mathbf{S}$ é uma função com flutuações estocásticas no processo de propagação que uniformemente faz escolhas aleatórias em cada simulação. Assim, a função de influência $\sigma(\mathbf{S})$ é estimada realizando um número suficiente de realizações da fração final de indivíduos informados $\varphi(\infty)\mathbf{S}$, com o conjunto de propagadores iniciais $\mathbf{S}$:

$$\sigma(\mathbf{S}) = \frac{1}{K} \sum_{\kappa=1}^K \varphi^{\kappa}(\infty)\mathbf{S} \quad (6.1)$$

onde $\varphi^{\kappa}(\infty)\mathbf{S}$ representa a fração final de indivíduos informados para uma determinada execução κ , e K é o número total de simulações para se obter uma boa estimativa do valor médio de $\sigma(\mathbf{S})$. Também calculamos a fração de propagadores ao longo do tempo $\phi\mathbf{S}(t) = |I(t)|/|V|$. Desta forma, a fração esperada de informantes em cada passo do tempo é definida por:

$$\phi\mathbf{S}(t) = \frac{1}{K} \sum_{\kappa=1}^K \phi^{\kappa}\mathbf{S}(t), \quad (6.2)$$

que é uma função côncava decrescente. Com esta formulação é possível calcular o passo do tempo esperado (t_p) no qual o maior pico de propagadores ocorre, onde $\{\exists t_p \geq 0 \mid \phi\mathbf{S}(t_p) - \phi\mathbf{S}(t) \geq 0, \forall t \geq 0\}$, e $\phi\mathbf{S}(t_p)$ é a fração máxima esperada de informantes durante a propagação.

A velocidade da propagação ($\bar{V}$) de um conjunto de propagadores iniciais $\mathbf{S}$ é definida como o ganho de indivíduos informados ao longo de um intervalo de tempo, isto é, $\bar{V} = \Delta \text{informados} / \Delta \text{tempo}$. Dessa forma, a velocidade para alcançar o pico de informantes $\bar{V}_p(\mathbf{S})$ está relacionada com a fração de novos propagadores atingida sobre o tempo gasto para atingir o pico, o qual é

$$\bar{V}_p(\mathbf{S}) = \frac{\phi\mathbf{S}(t_p) - \phi\mathbf{S}(0)}{t_p} \quad \text{para } t_p > 0 \quad (6.3)$$

onde $\phi\mathbf{S}(0) = |\mathbf{S}|/|V|$ é a fração do conjunto de sementes. Propomos esta nova medida para avaliar como a assortatividade da rede impacta nos métodos e na velocidade de propagação. Adoptamos $K = 600$ porque para este número de realizações os desvios-padrão são inferiores a 0.01 e o custo computacional de $\sigma(\mathbf{S})$ é pelo menos K vezes o custo polinomial associado

a $\varphi(\infty)^S$. Na verdade, o *IMP* é conhecido por ser um problema NP-hard para a maioria dos modelos de influência (KEMPE; KLEINBERG; TARDOS, 2015). As propostas para resolver o *IMP* são geralmente divididas em duas categorias (PEI; MAKSE, 2013; KEMPE; KLEINBERG; TARDOS, 2015):

1. Métodos heurísticos que analisam a topologia da rede, assumindo que existe uma forte relação entre a estrutura e o processo de propagação, mas sem garantia de resultados ótimos;
2. Métodos de otimização que analisam os modelos de influência e suas propriedades, assumindo que, por algumas estratégias de otimização, é possível fornecer uma garantia de aproximação de pelo menos 63% do valor ótimo, mas negligenciando a estrutura da rede e com um custo computacional considerável.

É importante enfatizar que os métodos de otimização podem alcançar soluções sub-ótimas, violando a garantia de aproximação, porque eles escalam mal com o tamanho do conjunto de sementes e da rede (DINH *et al.*, 2015), ou desconsideram a estrutura da rede e o nível de saturação da informação (GALSTYAN; MUSOYAN; COHEN, 2009).

A Figura 22 mostra os três métodos que consideramos para resolver o problema de maximização da influência: (i) selecionando os η vértices com os maiores valores de centralidade; (ii) detectando η comunidades na rede e, para cada comunidade isolada, selecionando o vértice com o maior valor de centralidade; e (iii) pelo método guloso de otimização que retorna os η melhores vértices propagadores. Os métodos recebem como entrada a rede G e o número requerido de η propagadores iniciais. O retorno é o conjunto $|S| = \eta$ sementes que maximizam a difusão de informação.

Métodos baseados na centralidade pressupõem que os vértices com maior centralidade devem convencer o maior número de indivíduos na rede (KITSACK *et al.*, 2010; KANDHWAY; KURI, 2016). O problema para esta heurística é descobrir qual centralidade identifica melhor os espalhadores mais influentes. O processo começa classificando os vértices da rede em ordem decrescente de centralidade. Depois disso, os η vértices mais centrais são os candidatos para o conjunto de sementes (Figura 22). Assim como em trabalhos anteriores (KITSACK *et al.*, 2010; HUANG *et al.*, 2014; KANDHWAY; KURI, 2016), aqui consideramos o grau (DG), centralidade de intermediação (BE) e PageRank (PR) para medir a centralidade de cada nó.

Outra abordagem heurística é baseada na estrutura das comunidades da rede. Consideramos a abordagem proposta de selecionar os propagadores iniciais dos vértices mais centrais dentro de cada comunidade. Verificou-se que as medidas de centralidade nas comunidades se comparam favoravelmente à escala global (HÉBERT-DUFRESNE *et al.*, 2013; VEGA-OLIVEROS; BERTON, 2015). Conforme ilustrado na Figura 22, as η principais comunidades da rede são detectadas e para cada comunidade G'_i , calculamos uma medida de centralidade e selecionamos

Figura 22 – Métodos aqui considerados para tratar do problema de maximização da influência (IMP).

Fonte: Adaptada de [Vega-Oliveros, Costa e Rodrigues \(2017a\)](#).

o seu vértice mais central. Portanto, obtemos η sementes nas quais a sobreposição de influências é minimizada, mas a influência dentro das comunidades é maximizada.

Também foi considerado o método de otimização baseado em um algoritmo guloso *hill-climbing* (KEMPE; KLEINBERG; TARDOS, 2003). Muitas abordagens têm sido derivadas da abordagem geral do método (KEMPE; KLEINBERG; TARDOS, 2015), onde a maioria tenta reduzir a complexidade computacional para alguma ordem polinomial (KEMPE; KLEINBERG; TARDOS, 2015; DINH *et al.*, 2015). No entanto, as propostas ainda são intratáveis no tempo para redes médias ou grandes, e os resultados podem ser iguais ou piores aos da abordagem geral (DINH *et al.*, 2015). Conscientes disso, analisamos apenas o método guloso geral, uma abordagem que garante um resultado de pelo menos 63% do caso ótimo (KEMPE; KLEINBERG; TARDOS, 2015). O algoritmo calcula entre todos os vértices $s_i \in V \setminus S$, isto é $\{s_i \in V \mid s_i \notin S\}$, aquele que maximiza a função $\sigma(S \cup \{s_i\})$, lembrando que S está inicialmente vazio. Depois, o

Figura 23 – Impacto da correlação grau-grau no problema de maximização de influência com $|\mathbf{S}|$ no intervalo de 2 a $90\%|V|$ de vértices. Para cada rede sem escala artificial calculamos o conjunto de sementes de acordo com: (BST) os vértices melhor classificados da rede; (COM) os vértices mais centrais das comunidades; (RANDOM) selecionando aleatoriamente as sementes; e (GREEDY) o método guloso de otimização. As medidas de centralidade adotadas são a centralidade de intermediação (BE), o grau (DG) e o PageRank (PR).

Fonte: Adaptada de [Vega-Oliveros, Costa e Rodrigues \(2017a\)](#).

vértice s_i é adicionado ao conjunto de sementes $\mathbf{S} = \mathbf{S} \cup \{s_i\}$ e o procedimento continua até que o conjunto de sementes tenha o tamanho $|\mathbf{S}| = \eta$.

Finalmente, definimos como solução média de referência a seleção uniforme de η vértices como sementes iniciais.

6.5 Resultados

6.5.1 Impacto da assortatividade em redes artificiais

Foi calculada a fração final de vértices contidos de acordo com a Equação 6.1. O número de sementes varia de 2 a $90\%|V|$ do número total de nós. O impacto da correlação de grau no problema de maximização da influência é mostrado na Figura 23. Observamos que ocorre um fenômeno inesperado quando as redes são disassortivas (veja Fig. 23 (a), (b) e (c)). Neste caso, a curva de nós influenciados ($\sigma(\mathbf{S})$) apresenta um pico quando o número de sementes são próximas a $10\%|V|$ da rede. Após essa porcentagem, a fração de informados diminui. Assim, o aumento no número de sementes pode não melhorar o alcance do rumor se espalhando em redes

AN. Esta redução de vértices influenciados deve-se ao aumento do número de contatos entre os propagadores, produzindo mais contidos em menos tempo. O número de vértices informados aumenta novamente quando o número de sementes é grande o suficiente para compensar esse efeito. Nas simulações, a fração de sementes necessárias para essa compensação aumenta com a disassortatividade das redes. Para redes assortivas, esta fração é de cerca de $30\%|V|$.

As curvas do método baseado em comunidades (COM) têm comportamento semelhante para todas as medidas de centralidade consideradas. O mesmo acontece com as curvas dos vértices melhores ranqueados ou classificados (BST). Assim, a seleção de diferentes medidas de centralidade não afeta significativamente a predição da fração de vértices influentes. Isso pode ser explicado pela correlação entre medidas de grau, PageRank e centralidade de intermediação nas redes BA. Assim, alguns vértices centrais tendem a fazer parte da mesma comunidade e a escolha de acordo com a comunidade é semelhante à escolha considerando toda a rede.

O pico em $\sigma(S)$ observado nas redes disassortivas é devido à baixa interação entre os informantes. Em uma rede disassortiva, os *hubs* estão conectados através de vértices de baixo grau, aumentando a área de interação entre eles, como mostrado na Figura 21. Se o número de sementes é maior que $10\%|V|$, a sobreposição de influência entre os propagadores é maior e eles se tornam contidos com mais frequência do que para um menor número de sementes. Para redes não correlacionadas e assortivas, selecionar os vértices centrais considerando as comunidades ou toda a rede leva ao menor valor $\sigma(S)$ e a influência da propagação é ainda pior do que a seleção uniforme de sementes (curvas RANDOM). Este efeito também se deve à interação entre as sementes informadas e informantes.

Comparando todos os casos e todo o intervalo de seleção de sementes, observamos que a propagação da informação é reforçada em redes com maior grau de correlação. No entanto, quando o número de sementes é menor a 10% da rede, a maior fração de vértices informados é obtido em redes disassortivas. Neste caso, a seleção é feita de acordo com a centralidade da rede sem escala. Assim, o alcance de um rumor depende do nível de assortatividade da rede, do método de seleção das semente e do número de sementes. Esse resultado não é esperado, pois o número de nós infectados deveria aumentar com o número de sementes. Este fenômeno é uma consequência da interação entre os vértices infectados, o que aumenta o número de contidos.

A Figura 24 mostra o comportamento da curva de vértices informados ao aumentar o número de sementes, de 2 a 10% do tamanho da rede, de acordo com a Equação 6.2. A medida que o número de semente é aumentado, o pico de vértices informados ocorre mais cedo. O maior pico de propagadores ocorre quando escolhemos as sementes de acordo com os vértices mais centrais nas redes disassortativas, superando os resultados do método de otimização guloso.

A velocidade para alcançar o pico de nós informados é mostrada na Figura 25. Somente em redes cujas sementes são selecionadas de acordo com o maior valor de centralidade da rede são afetados pelo nível de correlação grau-grau. Para os outros dois métodos (Fig. 25 (b) e (c)), a correlação de grau da rede não está aparentemente afetando a velocidade. Nestes casos, a

Figura 24 – Impacto da correlação grau-grau no pico de vértices informantes. O número de sementes está entre 2 a 10% do tamanho da rede (as setas indicam o aumento de η), selecionando as sementes pelo método guloso (*GREEDY*), comunidades (*COM*) e melhores ranqueados segundo a centralidade (*BST*).

Fonte: Adaptada de [Vega-Oliveros, Costa e Rodrigues \(2017a\)](#).

velocidade atinge um valor máximo e depois começa a decair lentamente. Esse resultado sugere que nas redes assortivas, as sementes mais bem classificadas tendem a interagir no início do processo de propagação, tornando-se rapidamente em contidos e diminuindo a velocidade de propagação.

6.5.2 Impacto da assortatividade em redes reais

A fração final de indivíduos informados (contidos) $\sigma(\mathbf{S})$ é calculada considerando-se o número de sementes no intervalo $[2, \eta_{\max}]$, onde $\eta_{\max}$ corresponde a $10\%|V|$. Os picos de influência são obtidos quando as sementes são selecionadas de acordo com medidas de centralidade (curvas de *BST*) ou comunidades (curvas de *COM*) (Fig. 26(a), (b) e (c)). Observe

Figura 25 – Velocidade do pico de vértices informantes no IMP de acordo com a correlação de grau. O número de sementes η está na faixa de 2 a $10\%|V|$ do tamanho da rede, selecionado pelos métodos: (a) melhores classificações (*BST*), (b) comunidades (*COM*) e (c) de otimização guloso *GREEDY*.

Fonte: Adaptada de [Vega-Oliveros, Costa e Rodrigues \(2017a\)](#).

que esses picos ocorrem somente em redes disassortivas. Este resultado é semelhante ao obtido em redes artificiais (ver Figura 23). Além disso, os picos são mais proeminentes quando as sementes são selecionadas de acordo com as medidas de centralidade (curvas *BST*). Depois de $|S| > 1\%|V|$, há uma diminuição rápida na fração de nós influenciados, à medida que a interação entre propagadores aumenta. Também observa-se na Figura 26 que a escolha das medidas de centralidade (*BE*, *DG*, *PR*) para selecionar as sementes em *BST* e *COM* tem um efeito insignificante sobre os valores de $\sigma(S)$. Isto também é observado para as redes artificiais (ver Figura 23).

A diferença entre os resultados observados para as redes artificiais e do mundo real pode ser devido à presença de triângulos e organização modular nas redes sociais (veja a Tabela 5). Contudo, o efeito da assortatividade é evidenciado em todos os casos.

O método de comunidades atinge valores de $\sigma(S)$ iguais ou superiores aos obtidos pela abordagem gulosa, em particular quando $|S| \leq 1\%|V|$. Isto sugere que para valores pequenos de $|S|$, as sementes extraídas das comunidade fornecem tantos vértices influenciados quanto os métodos de otimização gulosa. Isso se deve à alta organização modular das redes do mundo real. Como as redes artificiais não têm estrutura modular, a escolha das sementes segundo medidas de centralidade, considerando toda a rede ou as comunidades, tem efeitos semelhantes. No entanto, como em redes do mundo real as comunidades estão melhor definidas, os vértices centrais globais não correspondem aos vértices centrais locais dentro das comunidades. No entanto, devido ao baixo custo computacional envolvido na localização das sementes iniciais pelo método de comunidades, esta abordagem revelou ser a melhor para a maximização da influência em redes do mundo real.

Para as redes reais o comportamento de $\phi_S(t)$ é semelhante ao observado para as redes artificiais, com η menor que 10% de $|V|$. Além disso, o pico de propagadores $\phi_S(t_p)$ tende a diminuir com η apenas para o método *BST* em redes assortivas, mas aumenta em redes não

Figura 26 – Impacto da correlação de grau nas redes sociais. O número de sementes varia de 2 a 10 % do tamanho da rede, selecionado de acordo com as comunidades (*COM*), ranking dos nós mais centrais (best-rank, *BST*), aleatório (*RANDOM*) e pelo método de otimização gulosa (*GREEDY*). A centralidade de intermediação (*BE*), grau (*DG*) e PageRank (*PR*) são as medidas de centralidade utilizadas.

Fonte: Adaptada de Vega-Oliveros, Costa e Rodrigues (2017a).

Figura 27 – Velocidade de propagação para atingir o pico de propagadores. O número de sementes varia de 2 a 10% $|V|$, selecionando-os de acordo com os métodos: de otimização gulosa (*GREEDY*), comunidades (*COM*) e melhores classificados na rede (*BST*).

Fonte: Adaptada de [Vega-Oliveros, Costa e Rodrigues \(2017a\)](#).

correlacionadas e dissassortivas para todos os métodos. Em particular, em estruturas dissassortivas, o $\phi_S(t_p)$ de *BST* cresce com η até que a influência dos propagadores atinja o ponto de valor máximo $\sigma(S)$ e então começa a diminuir.

Na Figura 27 temos a velocidade para atingir o pico de informantes, $\bar{V}_p(S)$, para os três métodos e tamanhos diferentes de S , agrupados pelas redes: *Google+* (dissassortativa) (Fig. 27(a)), *email* (não correlacionado) (Fig. 27(b)); e *astrophysics* (assortiva) (Fig. 27(c)). Desta forma, podemos comparar $\bar{V}_p(S)$ entre os métodos e o tamanho de S . Notamos que o método *COM* apresenta velocidades de propagação mais altas do que a abordagem gulosa, para tamanhos de sementes iniciais inferiores a 2% do tamanho da rede. Isso sugere que o surto de informação é mais rápido ao selecionar as sementes iniciais pelas comunidades. Na Fig. 27(a), o método *BST* produz o valor de velocidade mais alto, em torno de 25% de novos propagadores novos por tempo, quando $|S|$ está no máximo valor de $\sigma(S)$. Os resultados de $\bar{V}_p(S)$ de *BST* tendem a cair na presença da assortatividade (Fig. 27(c)) como também verificado para as redes artificiais.

6.5.3 Distribuição das sementes

Espera-se que os rumores se espalhem melhor quando as sementes são distribuídas igualmente entre as comunidades. Isso deve ocorrer porque cada semente tenta infectar sua própria comunidade e a interação entre pares de propagadores é minimizada. Nós investigamos esta hipótese examinando a distribuição das sementes entre as comunidades. Nosso objetivo é quantificar as semelhanças entre os tamanhos das comunidade e a distribuição das sementes nas comunidades. Para esta tarefa aplicamos a variação normalizada da informação (*NVI*). Esta é uma métrica de Teoria da Informação que é normalizada em um sentido estocástico ([VINH; EPPS; BAILEY, 2010](#)). A medida baseia-se no ferramental de Teoria da Informação, definida a seguir: Dado dois conjuntos de variáveis discretas ou conjuntos $\mathbf{X}$ e $\mathbf{Y}$, sua entropia de informação conjunta ($\mathcal{H}$) e informações mútuas ($\mathcal{I}$) são expressas respectivamente em termos

Tabela 6 – A variação normalizada da informação (*NVI*) entre a distribuição de comunidades e as sementes GREEDY e BST-DG.

	Rede	S (%)	GREEDY <i>NVI</i>	BST-DG <i>NVI</i>
<i>DN</i>	<i>BA</i> [-0.43]	0.5	0.409	0.409
	<i>BA</i> [-0.43]	1	0.556	0.535
	<i>BA</i> [-0.43]	10	0.706	0.731
	<i>google+</i>	0.5	0.311	0.408
	<i>google+</i>	1	0.196	0.334
	<i>google+</i>	10	0.464	0.725
<i>NA</i>	<i>BA</i> [0.02]	0.5	0.409	0.409
	<i>BA</i> [0.02]	1	0.469	0.651
	<i>BA</i> [0.02]	10	0.708	0.843
	<i>email</i>	0.5	0.368	0.306
	<i>email</i>	1	0.524	0.695
	<i>email</i>	10	0.603	0.809
<i>AN</i>	<i>BA</i> [0.34]	0.5	0.344	0.689
	<i>BA</i> [0.34]	1	0.492	0.858
	<i>BA</i> [0.34]	10	0.693	0.861
	<i>astrophysics</i>	0.5	0.523	0.940
	<i>astrophysics</i>	1	0.589	0.948
	<i>astrophysics</i>	10	0.558	0.933

das distribuições marginais e conjuntas dos dados em $\mathbf{X}$ e $\mathbf{Y}$:

$$\mathcal{H}(\mathbf{X}, \mathbf{Y}) = - \sum_{\mathbf{x} \in \mathbf{X}} \sum_{\mathbf{y} \in \mathbf{Y}} p(\mathbf{x}, \mathbf{y}) \log p(\mathbf{x}, \mathbf{y}) \quad (6.4)$$

$$\mathcal{J}(\mathbf{X}, \mathbf{Y}) = \sum_{\mathbf{x} \in \mathbf{X}} \sum_{\mathbf{y} \in \mathbf{Y}} p(\mathbf{x}, \mathbf{y}) \log \left(\frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \right) \quad (6.5)$$

onde a probabilidade conjunta $p(\mathbf{x}, \mathbf{y})$ denota a probabilidade de que um item dos dados pertença a $\mathbf{X}$ e $\mathbf{Y}$. A informação mútua mede a sobreposição entre os dois conjuntos, porém não é uma métrica e nem está normalizada. Assim, para este estudo empregamos a *NVI* que é definida como

$$NVI(\mathbf{X}, \mathbf{Y}) = 1 - \frac{\mathcal{J}(\mathbf{X}, \mathbf{Y})}{\mathcal{H}(\mathbf{X}, \mathbf{Y})} \quad (6.6)$$

que está normalizada entre $[0, 1]$. A medida retorna o valor 0 quando os dois conjuntos são idênticos nas informações da distribuição dos seus itens e 1, quando há uma dissimilaridade completa entre as partições, ou seja, elas são independentes sem compartilhamento de informações.

Em termos da distribuição das sementes e das comunidades, identificamos η comunidades principais, onde cada comunidade é definida pelo subgrafo $G_c(V_c, E_c)$ e $\sum_{c=1}^{\eta} |G_c(V_c, E_c)| = |V|$. Como essas comunidades podem ter tamanhos diferentes, definimos $\mathbf{X}$ como o conjunto que

Tabela 7 – Resultados da influência ($\sigma(\mathbf{S})$) dos métodos analisados.

Rede	GREEDY		COM-DG		BST-DG	RANDOM
	$\sigma(1\%)$	$\sigma(10\%)$	$\sigma(1\%)$	$\sigma(10\%)$	$\sigma(1\%)$	$\sigma(1\%)$
<i>Google+</i>	0.1686	0.1663	0.1897	0.2627	0.1549	0.1501
<i>internet</i>	0.2130	0.2810	0.1906	0.2570	0.1639	0.1755
<i>caida</i>	0.1959	0.2640	0.1966	0.2531	0.1696	0.1743
<i>advogado</i>	0.4071	0.5179	0.4138	0.5002	0.3933	0.3821
<i>email</i>	0.6086	0.6976	0.6089	0.6707	0.5942	0.5941
<i>hamsterster</i>	0.5738	0.7318	0.5693	0.6871	0.5408	0.5236
<i>PGP</i>	0.3126	0.5356	0.3122	0.4985	0.2593	0.2654
<i>astrophysics</i>	0.5785	0.6979	0.5735	0.6745	0.5418	0.5456

descreve o tamanho das comunidades, onde a probabilidade de escolher um vértice e pertencer a alguma comunidade $\mathbf{x}$ é $p(\mathbf{x}) = |V_{\mathbf{x}}|/N$. Por outro lado, dizemos que $\mathbf{Y}$ é o conjunto que denota o número de sementes nas comunidades, com $p(\mathbf{y}) = |\mathbf{S}^{\mathbf{y}}|/\eta$ a probabilidade de que uma semente pertença a alguma comunidade $\mathbf{y}$.

Calculamos a similaridade NVI para as sementes do método guloso e as melhores classificadas da rede, que é apresentado na Tabela 6. Quanto mais igualmente distribuídos o conjunto de sementes com o tamanho das comunidades, o valor NVI tende a 0. Quando as sementes estão concentradas em apenas uma comunidade, os valores de NVI tendem a 1. Valores pequenos de NVI significam que as sementes estão distribuídas de forma mais homogênea segundo o tamanho das comunidades. Notamos na Tabela 6 que as sementes *GREEDY* possuem valores NVI inferiores aos das sementes *BST-DG*. Isso significa que as sementes do método guloso estão mais homogêneamente distribuídas entre as comunidades. Nas redes artificiais, para tamanhos de $|\mathbf{S}| = 0.5\%|V|$, as sementes tendem a estar mais igualmente concentradas nas comunidades. No entanto, para tamanhos maiores de $\mathbf{S}$, as sementes concentram-se mais em comunidades ou regiões específicas. Podemos ver que os resultados de NVI do método *BST-DG* são maiores do que *GREEDY*, o que indica que as sementes *BST-DG* tendem a estar mais concentradas em uma região específica da rede. Por outro lado, nas redes reais, as sementes do método guloso estão distribuídas mais regularmente nas comunidades do que as sementes mais centrais da rede (observe que os valores de NVI para *GREEDY* estão abaixo de 0.6). Este resultado sustenta a hipótese de que as sementes do método guloso tende a se distribuir seguindo a estrutura de comunidades da rede. Isso também indica que o *IMP* pode ser abordado como o problema de identificar a melhor seleção de sementes iniciais segundo à distribuição das comunidade.

Figura 28 – A diferença crítica (CD) para comparar a classificação média de métodos diferentes no 95 percentil é 1.66, de acordo com o teste de Nemenyi. Diferenças significativas acima deste valor são significativas e estão desconectadas.

Fonte: Vega-Oliveros, Costa e Rodrigues (2017a).

6.6 Análise estatística

Nesta seção abordamos a questão de como comparar os resultados dos múltiplos métodos nas múltiplas redes e procurar diferenças significativas entre os algoritmos. Inicialmente, analisamos os resultados da maximização da influência dos métodos considerando o tamanho de S igual 1% de $|V|$ (Tabela 7). Para este propósito, realizamos um teste de significância estatística empregando a abordagem de Friedman e Nemenyi (DEMSAR, 2006). O teste de Friedman é uma contraparte não-paramétrica da bem conhecida ANOVA, com o correspondente teste post-hoc de Nemenyi para comparação das médias dos algoritmos. Se a hipótese nula de desempenho semelhante é rejeitada pelo teste de Friedman, prosseguimos com o teste post-hoc de Nemenyi para comparações por pares, verificando se as diferenças nos valores de classificação são estatisticamente significativas.

Para esta análise em particular adotamos a representação do diagrama crítico sugerida por Demšar (DEMSAR, 2006). No diagrama, uma linha horizontal representa o eixo no qual os valores de classificação médios dos métodos são plotados. No eixo, os métodos melhores classificados estão no lado esquerdo. Algoritmos que não são significativamente diferentes uns dos outros estão conectados através de uma linha horizontal. O desempenho entre os métodos é significativamente diferente se suas posições de classificação correspondentes diferirem pelo menos pela diferença crítica (CD). O resultado do CD do teste de Nemenyi é apresentado no topo do diagrama.

De acordo com o resultado de $\sigma(1\%)$ da Tabela 7, as estatísticas Chi-Quadrado para os métodos são 19,20 e o valor crítico das estatísticas Chi-quadrado com 3 Graus de liberdade no 95 percentil é 7,81. Para o teste de Friedman usando a estatística Chi-Quadrado, a hipótese nula de que todos os métodos se comportam de forma semelhante deve ser rejeitada. Além disso, calculamos também a F-estatística dos métodos, que resultou em 28.00. Com 3 e 21 graus de liberdade e no 95 percentil, o valor crítico da estatística F é 3,07. Isto significa, de acordo com o teste de Friedman, que a hipótese nula em que todos os métodos se comportam de forma semelhante deve ser rejeitada.

Desse modo, aplicamos o teste post-hoc de Nemenyi para descobrir qual método obtém melhores resultados. O diagrama crítico do teste de Nemenyi é mostrado na Figura 28. O CD para comparar o ranking médio entre dois métodos no 95 percentil é 1,66. As diferenças de médias de rankings acima deste valor são significativas. Encontramos que não há diferença estatisticamente significativa nos resultados de maximização de influência entre o método *GREEDY* e *COM*, para tamanhos de sementes iniciais até $1\%|V|$. No entanto, o teste de Nemenyi indica diferença estatística significativa entre as abordagens gulosa e comunitária com os métodos aleatórios e vértices mais centrais da rede.

Os resultados de $\sigma(\mathbf{S})$ para *BST* e *RANDOM* tornam-se piores do que *GREEDY* e *COM* para $|\mathbf{S}| > 1\%|V|$ nas redes do mundo real (Figura 26). Portanto, calculamos a significância estatística apenas com os métodos guloso e de comunidades. Para fazer a avaliação de dois algoritmos em conjuntos de dados múltiplos, empregamos o recomendado teste estatístico de Wilcoxon (DEMSAR, 2006). Este teste estatístico é uma alternativa não-paramétrica ao teste t-test. Também, o teste de Wilcoxon é mais sensível, os *outliers* têm menor efeito e não assumem distribuições normais do que o no caso do t-test (DEMSAR, 2006). Para rejeitar a hipótese nula de desempenho semelhante, o W -value retornado pelo teste deve ser menor do que o correspondente valor crítico W_c da tabela de teste Wilcoxon.

Foi calculado o teste estatístico de Wilcoxon no 95 percentil para as abordagens gulosa e comunitária, com os resultados $\sigma(\mathbf{S})$ das redes do mundo real considerando $|\mathbf{S}| = 10\%|V|$ (Table 7). Como resultado, obtemos um W -value = 8. O valor crítico para oito redes em $p = 0,05$ da um valor crítico de $W_c = 3$. Portanto, o resultado não é significativo, o que significa que a hipótese nula de desempenho semelhante dos métodos não pode ser rejeitada. O resultado sugere que a maximização da influência entre os métodos guloso e baseado em comunidades até $|\mathbf{S}| = 10\%|V|$ não são estatisticamente diferentes.

Neste ponto, chamamos a atenção ao custo computacional que o método de otimização gulosa e derivações possui. Ainda que teóricamente apresentam garantia de aproximação $(1 - 1/e)$ ($\approx 63\%$ do ótimo), tem um custo computacional muito alto, estando próximo a NP-hard (KEMPE; KLEINBERG; TARDOS, 2015; DINH *et al.*, 2015). Por outro lado, os métodos heurísticos são muitas ordens de magnitude mais rápidos do que as abordagens da família hill-climbing para o IMP. No entanto, aqui mostramos que os métodos heurísticos, como os vértices mais centrais nas redes disassortivas ou as sementes mais importantes das comunidades, também produzem conjuntos de sementes de alta qualidade mesmo sem prova teórica de garantia de aproximação. Além disso, a abordagem por comunidades atinge o pico do informantes mais rapidamente, e na prática, obtém resultados com desempenho estatístico semelhante ao do método guloso.

6.7 Considerações finais

Neste capítulo analisamos o impacto da assortatividade no problema da maximização da influência em redes complexas, sob o modelo de rumor *SIR* de [Maki e Thompson \(1973\)](#). Foi variado o tamanho das sementes iniciais de 2 até 10% do tamanho das redes do mundo real, e até 90% para as redes artificiais. Por exemplo, nas redes disassortativas, a seleção dos melhores classificados, ou mais centrais, como vértices sementes produz os resultados mais destacados de influência, crescendo proporcionalmente com o conjunto de sementes até um tamanho máximo de saturação. Isso pode ser entendido porque vértices importantes, como *hubs*, estão mais distantes uns dos outros. Portanto, suas influências não se sobrepõem nas vizinhanças e o potencial de propagação não é prejudicado até o tamanho de número máximo de sementes. Consequentemente, para redes assortativas, a seleção dos vértices mais centrais leva a uma menor propagação, mesmo pior que a seleção aleatória de sementes.

Estratégias de seleção das sementes iniciais que não consideram a estrutura da rede, como as abordagens de otimização gulosa, podem produzir resultados subótimos. Assim, foi proposto um método escalonável e robusto baseado na seleção de sementes de acordo com a estrutura de comunidades da rede. Esta abordagem mantém um bom alcance da influência comparável ao algoritmo do estado da arte, mas sendo muito mais rápido. Foram feitas extensas simulações em redes artificiais e do mundo real demonstrando que nosso método heurístico tem influência correspondente com um algoritmo de aproximação gulosa padrão, sem diferenças estatísticas significativas. Uma vantagem adicional da abordagem de comunidades é que para conjuntos de sementes iniciais inferiores a $1\%|V|$, a velocidade dos surtos é maior do que a abordagem de otimização gulosa, indicando que o pico dos propagadores é atingido em menor tempo.

Observamos também que a distância entre as sementes tem pouca influência sobre o processo de propagação. Na verdade, o método guloso, de forma implícita, seleciona as sementes para que elas fiquem distribuídas entre as comunidades da rede. Isso indica que o *IMP* pode ser reduzido ao problema de identificar a melhor seleção de sementes iniciais quanto à distribuição das comunidades. Também indica que o problema de maximização da influência pode ser abordado como a identificação do melhor conjunto de propagadores iniciais de acordo com as comunidades da rede.

Além disso, as estruturas de dados simples e a computação iterativa do método proposto podem ser aplicadas a implementações paralelas e distribuídas, para ampliar ainda mais as análises de maximização de influência em redes sociais muito grandes (milhões ou bilhões de vértices).

Dados os resultados na influência estatisticamente semelhantes, a maior velocidade de propagação e o baixo custo computacional do nosso método ($\mathcal{O}(n \log n)$) em relação à abordagem de otimização gulosa (*NP*-difícil), a seleção de vértices influentes por comunidades é uma área de pesquisa promissora para o problema de maximização da influência e outros em redes complexas.

IDENTIFICAÇÃO DOS PROPAGADORES MENOS INFLUENTES DA REDE

Com o surgimento das mídias sociais aumentou-se a necessidade de reconhecimento da influência social, motivados principalmente pela publicidade online, campanhas políticas e de saúde, sistemas de recomendação, estudos epidemiológicos, entre outros. Nos processos de propagação é possível identificar os vértices mais centrais ou influentes de acordo com a topologia e a dinâmica da rede. Por outro lado, a identificação dos propagadores menos influentes tem sido desconsiderada. Este capítulo visa definir e identificar a presença de vértices menos influentes na propagação de rumores. Junto a isto, propomos uma estratégia adaptativa para tornar os piores propagadores em melhores espalhadores e maximizar a média de propagação de informação na rede. Os resultados experimentais em três redes sociais (*Google+*, *hamsterster* e *advogato*) confirmam que selecionando 0.5% dos propagadores menos influentes e trocando uma conexão para algum vértice importante, aumentamos em média a capacidade de propagação desses vértices e de toda a rede.

O capítulo está estruturado como segue: na Seção 7.1 apresentamos a motivação de identificar os vértices menos influentes da rede. Na Seção 7.2 apresentamos a definição dos vértices que são os vértices menos influentes na propagação de informação. As redes complexas consideradas e os métodos empregados para a identificação dos propagadores menos influentes são apresentados na Seção 7.3 e Seção 7.4 respectivamente. A proposta de um método adaptativo para melhor a capacidade de propagação desses vértices e os resultados são apresentados na Seção 7.5 e Seção 7.6 respectivamente. Finalmente, as considerações finais do capítulo estão na Seção 7.7.

7.1 Motivação

Atualmente as mídias sociais, como blogs, redes sociais, sites de compartilhamento, entre outros, podem espalhar rumores ou informações que atinjam milhões de usuários em um curto intervalo de tempo. Processos de propagação de informações relacionados com rumores podem ser benéficos ou destrutivos em cenários como mercado de ações, publicidade e marketing, adoção de tecnologias, política e segurança nacional, entre outros.

Desde o estabelecimento da área de redes complexas, a teoria de redes fornece uma modelagem mais realista e precisa de processos de propagação em sistemas complexos. A representação desses sistemas complexos como redes possibilita a inclusão de sua estrutura heterogênea nos modelos tradicionais, o que tem forte influência na disseminação de informações (MORENO; NEKOVEE; PACHECO, 2004; BARRAT; BARTHÉLEMY; VESPIGNANI, 2008).

Assim, os estudos em propagação de informação ou rumores concentram-se em como as idéias são compartilhadas entre os indivíduos e quais são as condições que permitem uma disseminação massiva (MORENO; NEKOVEE; PACHECO, 2004; CASTELLANO; FORTUNATO; LORETO, 2009; GONZÁLEZ-BAILÓN; Borge-Holthoefer, Javier Rivero; MORENO, 2011; BORGE-HOLTHOEFER *et al.*, 2012). Nesse sentido, é possível encontrar os vértices mais centrais ou influentes durante o processo de difusão. Por exemplo, alguns indivíduos podem ter uma influência social mais alta do que outros, de acordo com sua posição e características topológicas na rede (KITSAK *et al.*, 2010; GONZÁLEZ-BAILÓN; Borge-Holthoefer, Javier Rivero; MORENO, 2011; BORGE-HOLTHOEFER *et al.*, 2012). Muitas pesquisas buscam identificar os melhores propagadores em diferentes tipos de redes (BURT, 1992; SABIDUSSI, 1996; KITSAK *et al.*, 2010; BORGE-HOLTHOEFER *et al.*, 2012). Por outro lado, a identificação dos vértices menos influentes na rede tem sido ignorada.

Consideramos que os propagadores menos influentes podem ser importantes no processo de propagação, pois eles podem representar grupos ou regiões densas de indivíduos dentro da rede. Por exemplo, considere o cenário de uma rede de colaboração científica, onde alguns institutos de pesquisa são mais proeminentes do que outros. Se um pesquisador com poucas publicações ou contatos visita um grupo proeminente, pelo fenômeno de homofilia pode, consequentemente, melhorar sua capacidade de publicação e impacto na rede.

Desse modo, a identificação de usuários menos influentes pode ser de grande utilidade em diferentes processos de propagação. Métodos que identifiquem vértices pouco influentes mas que tenham um grande potencial para se tornar melhores propagadores ou aumentar o fluxo de informação na rede tem sido pouco explorado. Também, verificar como a estrutura pode influenciar a escolha desses propagadores e a proposta de mecanismos simples e de baixo custo para aumentar consideravelmente a capacidade de propagação desses vértices e da rede como um todo se torna um novo desafio na área.

Tabela 8 – Propiedades estruturais das redes complexas

Rede	N	$\langle DG \rangle$	$\langle CT \rangle$	$\langle HO \rangle$	$\langle BE \rangle$	$\langle CL \rangle$	$\langle EV \rangle$	$\langle KC \rangle$	$\langle PR \rangle$
<i>hamsterster</i>	2000	16.1	0.5399	0.2898	2.588×10^3	1.43×10^{-4}	0.011	9.287	5.00×10^{-4}
<i>advogado</i>	5054	15.6	0.2525	0.3700	5.747×10^3	6.19×10^{-5}	6.82×10^{-3}	8.137	1.98×10^{-4}
<i>google+</i>	23613	3.32	0.1742	0.8112	3.581×10^4	1.10×10^{-5}	2.30×10^{-3}	1.669	4.21×10^{-5}

7.2 Definição dos vértices menos influentes

Analizamos a capacidade de propagação nas escalas micro (vértices) e macro (rede). Para cada vértice $i \in V$, calculamos a fração final de informados por nó, isto é $\varphi(\infty)^i$ onde i é o vértice propagador inicial ou semente. Esta medida quantifica o alcance da propagação do rumor a partir de i . Para obter os valores da medida foi feita uma média entre 90 realizações para cada vértice.

Na escala macroscópica, $\varphi(\infty)^V$ representa a média da capacidade de propagação de todos os vértices da rede, o qual significa

$$\varphi(\infty)^V = \frac{\sum_{i \in V} \varphi(\infty)^i}{N}. \quad (7.1)$$

Sendo assim, definimos que os vértices menos influentes da rede são aqueles cuja capacidade de propagação é menor que a metade da média de propagação da rede, isto é $\varphi(\infty)^i \leq \frac{\varphi(\infty)^V}{2}$.

7.3 Redes consideradas

As seguintes redes sociais foram analisadas neste trabalho: *hamsterster* (KUNEGIS, 2016), *advogado* (ADVOGATO..., 2016) e a rede *Google+* (GOOGLE+..., 2016), assumindo as redes como não direcionadas e considerando o componente principal. As propriedades topológicas das redes sociais selecionadas estão sumarizadas na Tabela 8, com as médias das seguintes medidas de centralidade: o grau $\langle DG \rangle$, coeficiente de aglomeração $\langle CT \rangle$, buracos estruturais $\langle HO \rangle$, grau de intermediação $\langle BE \rangle$, grau de proximidade $\langle CL \rangle$, grau de autovetor $\langle EV \rangle$, k -core $\langle KC \rangle$ e PageRank $\langle PR \rangle$.

7.4 Os propagadores menos influentes

Para avaliar o impacto dos propagadores menos influentes na rede, empregamos a medida z-score sobre as capacidades de propagação $\varphi(\infty)^i$ e classificamos os valores em ordem crescente. O z-score indica quantos desvios padrão um elemento está da média dos valores. Para um elemento x , o z-score é definido como

$$z = \frac{x - \langle x \rangle}{\sigma}, \quad (7.2)$$

Figura 29 – Comparativo de z-score da capacidade de propagação em função da proporção de propagadores menos influentes.

Fonte: Adaptada de [Vega-Oliveros et al. \(2015\)](#).

onde $\langle x \rangle$ é a média e σ o desvio padrão das amostras. O impacto é apresentado na Figura 29. Considerando 10% dos vértices com menor capacidade de propagação, a média tem um decaimento de cerca de 3 vezes o valor desvio padrão. Este comportamento decrescente continua sucessivamente, assim, evidenciamos que existem vértices propagadores menos influentes e que estes influenciam de forma negativa na média global da capacidade de propagação da rede. Nas simulações foi adotado o *CP*, como regra de comportamento dos propagadores.

7.4.1 Análise topológica

Calculamos os valores de z-scores para diferentes medidas de centralidade dos vértices. As medidas locais utilizadas para a caracterização dos vértices foram: centralidade de grau (*DG*), centralidade de intermediação (*BE*), centralidade de PageRank (*PR*), centralidade de proximidade (*CL*), centralidade de *K*-core (*KC*), centralidade de autovetor (*EV*), centralidade de aglomeração (*CT*), e a centralidade de buracos estruturais (*HO*). O resultado do z-score para cada centralidade e rede social é apresentado na Figura 30(a-c). A fração de indivíduos foi ordenada de modo crescente para cada centralidade, onde os vértices com os menores valores de centralidade estão a esquerda. Observamos nas três redes sociais que os vértices com mais baixa centralidade *EV* são aqueles que produzem os menores valores de z-score.

Empregamos o coeficiente de Jaccard para entender melhor a relação entre as medidas de centralidade e a capacidade de propagação dos vértices. O coeficiente de Jaccard compara a diversidade entre conjuntos de amostras. O resultado da medida é o cálculo da similaridade entre dois conjuntos de amostras finitas $\mathcal{A}$ e $\mathcal{B}$, definido como o tamanho da interseção dividido pelo

Figura 30 – Análise de propagação para as redes sociais reais *hamsterster*, *advogato* e *Google+*: (a-c) o impacto de vértices com baixa centralidade no processo de propagação; (d-f) similaridade entre os vértices com baixa centralidade e os menos influentes; (g-i) características topológicas dos propagadores menos influentes e os vértices com os menores valores de *EV*. As simulações para *hamsterster* e *advogato* foram considerados os valores de $\beta = 0.3$ e $\mu = 0.1$, e para *Google+* $\beta = 0.5$ e $\mu = 0.1$.

Fonte: Adaptada de [Vega-Oliveros et al. \(2015\)](#).

tamanho da união dos conjuntos

$$J(\mathcal{A}, \mathcal{B}) = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|} \quad (7.3)$$

Assim, usamos o coeficiente de Jaccard para medir a proporção de propagadores menos influentes dentro do conjunto de vértices com baixa centralidade, onde $\mathcal{A}$ representa a fração dos vértices segundo uma medida de centralidade e $\mathcal{B}$ é o conjunto de vértices com menor capacidade de propagação. (Figura 30(d-f)). Podemos observar que ao selecionar até 1% dos vértices com menores valores de centralidade de autovetor, cerca de metade desses vértices são propagadores menos influentes.

Tabela 9 – Resultados do método de agrupamento *k*-means para a rede *Google+*

Medidas	Grupo1 (10%)	Grupo2 (67%)	Grupo3 (23%)
$\langle DG \rangle$	1.029 ± 1.001	1.128 ± 0.386	10.553 ± 72.435
$\langle BE \rangle$	0.00 ± 0.0004	0.00 ± 0.0005	0.0023 ± 0.023
$\langle CL \rangle$	0.5 ± 0.00	0.5 ± 0.00	0.501 ± 0.019
$\langle PR \rangle$	0.0007 ± 0.0007	0.0007 ± 0.0002	0.0043 ± 0.0299
$\langle EV \rangle$	0.00 ± 0.00	0.0021 ± 0.0037	0.0117 ± 0.0202
$\langle KC \rangle$	0.083 ± 0	0.094 ± 0.032	0.292 ± 0.178
$\langle CT \rangle$	0.00 ± 0.00	0.00 ± 0.00	0.746 ± 0.312
$\langle HO \rangle$	0.991 ± 0.0291	0.942 ± 0.166	0.360 ± 0.145
Difusão	2612.274 ± 332.095	3112.858 ± 56.965	3135.407 ± 37.310

Consecutivamente, analisamos as propriedades topológicas de 1% dos vértices menos influentes e com menor valor de *EV*. Verificamos se esses vértices também possuem valores baixos em outras centralidades. Os valores de centralidade foram normalizados pela média e foi graficado o maior valor registrado para cada conjunto de vértices (Figura 30(g-i)). Observamos que as propriedades topológicas dos vértices com os menores valores de *EV* estão contidas dentre os valores máximo do grupo de propagadores menos influentes. Diferentemente do esperado, vértices com valores de centralidade acima da média como no caso de *BE* (Figura 30(h,i)), *DG* (Figura 30(i)) ou *PR* (Figura 30(g-i)) podem também ser propagadores menos influentes.

7.4.2 Análise não supervisionada

Também fizemos uma análise não supervisionada visando reconhecer o grupo de propagadores menos influentes de acordo com as características topológicas dos vértices e sua capacidade de propagação. Temos como hipótese que objetos do mesmo grupo tendem a ser mais similares entre si que comparados com objetos de outros grupos. Portanto, buscamos identificar as propriedades que têm em comum os vértices que pertencem ao grupo dos piores propagadores.

Para isso aplicamos o método de agrupamento *k*-means (MACQUEEN, 1967), o qual particiona os dados em *k* grupos mutuamente exclusivos. Cada grupo é definido por um centróide, o qual representa o ponto em que a soma das distâncias para todos seus elementos é minimizada. Usamos a versão do algoritmo desenvolvido pelo Weka¹, com as configurações padrões.

Realizamos a tarefa de detecção dos grupos nas redes *hamsterter*, *advogado* e *Google+*. Verificamos que em todos os casos os vértices com centralidade *EV* perto de zero estão no grupo dos propagadores menos influentes. Por exemplo, os resultados para a rede *Google+* são mostrados na Table 9. Nesse caso, o parâmetro do *k*-means (que indica o número de grupos) foi definido em 3 e por isso que a rede está dividida em três grupos. Estamos interessados no Grupo1, dado que possui os vértices com a menor média de capacidade de propagação. Ao compararmos as médias das medidas de centralidade entre os grupos identificados, vemos que os

¹ <http://www.cs.waikato.ac.nz/ml/weka>

valores de $\langle EV \rangle \approx 0$ e $\langle HO \rangle \approx 1$ se destacam no Grupo 1. No caso, para as redes *hamsterter* e *advogado*, as médias das centralidades que se destacaram no grupo dos que menos difundem a informação foram $\langle CL \rangle < 0.5$ e também $\langle EV \rangle \approx 0$.

Portanto, o método não supervisionado que usamos para encontrar o conjunto dos vértices que menos propagam na rede consiste em extrair as medidas de centralidade dos vértices, executar o algoritmo k -means até chegar a uma divisão de grupos em que algum deles atinjam um valor de $\langle EV \rangle < 1/N$. Esse padrão foi verificado de modo experimental com vários tipos de redes. Isso também está em concordância com os resultados da análise topológica. Desse modo, podemos generalizar, sem precisar de calcular a capacidade de propagação de cada vértice $\varphi(\infty)^i$, que os vértices com os menores valores de EV , tendem a ter uma menor capacidade de propagação na rede.

7.5 Método adaptativo para melhorar a propagação de informação

Uma abordagem é considerada adaptativa quando existe uma evolução da dinâmica com a respectiva adaptação da topologia e vice versa (BORNHOLDT; ROHLF, 2000; GROSS; CARLOS; BERND, 2006; GROSS; BLASIUS, 2008; SAYAMA *et al.*, 2013). Nesse sentido, a evolução acontece como os estados dos vértices, parâmetros da dinâmica ou a estrutura da rede, o qual pode gerar uma interação mútua e complexa que varia no tempo. As pesquisas baseadas em redes adaptativas apresentam duas principais abordagens: (i) consideram que a rede possui um número fixo de vértices cujas conexões se adaptam dependendo das condições da dinâmica e os estados dos vértices (READ; KEELING, 2003; GROSS; CARLOS; BERND, 2006). Assim, a adaptação está baseada no processo de reconexão dos vértices considerando novas condições. (ii) consideram a complexidade do crescimento da rede enquanto acontece a dinâmica (GUERRA; NES; LATROA, 2012). Na adaptação da rede, as novas conexões para os vértices inseridos dependem do estado dos vértices que já formam parte dela, seguindo as condições da dinâmica. Porém, uma vez que as conexões forem feitas estas não são alteradas.

Estamos interessados na primeira abordagem, ou seja, adaptação da rede a partir da troca de conexões. Este tipo de adaptação pode ser vista como o caso da propagação de uma doença na sociedade. Quando surge uma nova epidemia, de modo intuitivo as pessoas evitam se expor e entrar em contato com indivíduos doentes. Assim mesmo, as pessoas doentes tendem a ficar em centros médicos ou em casa para se recuperarem. Essa simples conduta pode levar a diferentes efeitos na dinâmica da doença, influenciada pelo processo de mudança de contatos das pessoas. Gross, Carlos e Bernd (2006) propuseram no modelo SIS em que os indivíduos suscetíveis tentem evitar o contato com os infectados. No modelo, um indivíduo suscetível deixa de se conectar com um vizinho infectado com uma probabilidade w , e passa a se conectar com outro suscetível. Com essa estratégia, a estrutura da rede é fortemente afetada, passando a ter

uma interação muito próxima com a dinâmica e podendo ser fragmentada em componentes isolados, o qual não é desejável para as redes sociais. Liu e Zhang (2014) propõem uma abordagem adaptativa considerando o modelo *SIR* reativo, a qual recombina as conexões dos vértices informantes selecionando os vizinhos que lhe ofereçam uma maior recompensa segundo uma função de Fermi (SANTOS; PACHECO; LENAERTS, 2006; PACHECO; TRAULSEN; NOWAK, 2006). Na abordagem adaptativa proposta, o vértice informante rompe a conexão com um ignorante quando a recompensa é menor à hipótese de conectar-se aleatoriamente com algum dos seus segundos vizinhos ignorantes. Os autores encontraram que a informação propagou mais rapidamente em comparação com os resultados das versões estáticas das redes. No entanto, os resultados do processo adaptativo dependem das condições iniciais da simulação, como a seleção das sementes, definindo assim o alcance e a velocidade de difusão da informação.

Neste projeto objetivamos um método adaptativo de baixo custo, onde ao trocar uma pequena fração de vértices e arestas seja possível melhorar o alcance da difusão de informação, sem afetar expressivamente a estrutura da rede. Para isso, exploramos e identificamos os propagadores menos influentes da rede, considerando medidas de centralidade e abordagens topológicas e de agrupamento de dados. Os resultados indicam que os vértices com centralidade de autovetor (*EV*) com valor $X_i < 1/N$ têm pouca influência no processo de propagação. Assim, após reconhecer estes propagadores menos influentes, buscamos melhorar sua capacidade de difusão.

O método adaptativo consiste em selecionar uma fração dos piores propagadores e, em seguida, para cada um deles fazer a troca aleatória de apenas uma aresta para um vértice central, também escolhido uniformemente. Consideramos como vértices centrais ou influentes aqueles com maior medida de centralidade na rede. Com esta abordagem podemos melhorar tanto a capacidade de propagação dos propagadores menos influentes quanto a capacidade global de difusão, sem modificar a distribuição de grau nem o número de arestas da rede. Outra vantagem do método é que não depende de parâmetros ou condições iniciais, sendo este generalizável. A proposta baseia-se na ideia que ao selecionar alguns poucos propagadores não influentes e apresentá-los a um vértice influente, podemos melhorar sua capacidade tornando-os melhores propagadores e, conseqüentemente, incrementar o alcance global da difusão de informação.

7.6 Resultados do método adaptativo

Para melhorar a capacidade de propagação da rede, selecionamos 5%, 2.5% e 0.5% dos vértices com menores valores de *EV*. Para cada vértice apenas uma de suas arestas foi selecionada aleatoriamente e reconectada a um vértice influente da rede. Consideramos como vértices influentes aqueles com maior centralidade *DG*, *BE*, *EV* e *PR*. Para cada vértice influente selecionado trocamos apenas uma aresta em todo o processo.

Para cada medida de centralidade e fração de vértices com baixo *EV*, calculamos a

Tabela 10 – Melhoria na capacidade global de propagação para as redes sociais: Primeiros lugares estão em negrito e segundos lugares estão sublinhados.

Rede	Arestas trocadas (%)	DG	BE	EV	PR
<i>hamsterster</i>	5	1.0817	1.0932	1.0810	<u>1.0844</u>
	2.5	<u>1.0804</u>	1.0817	1.0794	1.0783
	0.5	<u>1.0493</u>	1.0510	1.0479	1.0489
<i>advogato</i>	5	1.0600	1.0594	<u>1.0599</u>	1.0594
	2.5	1.0529	<u>1.0556</u>	1.0549	1.0557
	0.5	<u>1.0450</u>	1.0439	1.0453	1.0447
<i>google+</i>	5	<u>1.0124</u>	1.0158	1.0050	1.0112
	2.5	<u>1.0143</u>	1.0102	1.0103	1.0171
	0.5	<u>1.0241</u>	1.0190	1.0187	1.0266

Figura 31 – Análise de propagação para as redes sociais ao realizar a proposta adaptativa trocando uma aresta dos 0.5% dos propagadores menos influentes, e conectando-os com os vértices com maior grau da rede: o impacto dos valores de baixa centralidade na média de propagação.

Fonte: Adaptada de [Vega-Oliveros et al. \(2015\)](#).

capacidade de propagação global ($\varphi(\infty)^V$) normalizada pela média do caso padrão ou sem mecanismo de adaptação (Tabela 10). Ao conectar os propagadores menos influentes com os *hubs* da rede obteve-se 6 segundos lugares e 1 primeiro lugar nos resultados. O caso de selecionarmos 0.5% dos vértices com menor valor de *EV* e conectando-os aos *hubs* da rede é ilustrado na Figura 31. Observamos na tabela que existe pouca diferenças entre 5% e 2.5%, mas entre 2.5% e 0.5% a diferença aumenta. Também, temos o fator de custos computacional para calcular as diferentes medidas, sendo muito mais direto calcular a centralidade *DG* que *BE*. Verificamos que a média da capacidade de propagação aumentou para as redes, os valores de z-score aumentaram não apenas para o 0.5% dos vértices selecionados mas também nas outras medidas e na rede toda (Figura 31). Além disso, para os melhores casos, a média das capacidades de propagação da rede do método adaptativo foi maior que a capacidade de propagação do melhor e mais influente vértice da rede no estado convencional.

Assim, a abordagem proposta melhorou a média da capacidade de propagação das três redes analisadas: *Google+*, *hamsterster* e *advogato*. Esses resultados indicam que existe um grande potencial nas técnicas adaptativas que considerem os vértices menos influentes, o qual

tem sido pouco explorado na literatura. Melhores e mais eficientes métodos de re-conexão podem ser desenvolvidos para incrementar o fluxo de informação ou de um sinal nas redes complexas.

7.7 Considerações finais

Neste trabalho identificamos e confirmamos a existência dos propagadores menos influentes no processo de propagação de rumores. Caracterizamos os propagadores menos influentes analisando medidas de centralidade, tais como grau, intermediação, proximidade, PageRank, autovetor, k -core, coeficiente de agrupamento e buracos estruturais. Seleccionamos uma proporção de vértices menos influentes das redes sociais *Google+*, *hamsterster* e *advogato* e os tornamos melhores espalhadores, reconectando uma de suas arestas com um vértice mais central. Depois disso, analisamos o impacto com diferentes proporções no processo adaptativo e observamos que a capacidade global de propagação de informação pode aumentar em até um 5% trocando apenas uma única aresta de 0.5% dos vértices da rede.

Esta é uma pesquisa exploratória e em andamento que apresenta resultados experimentais promissores. Novas propostas de estudo podem ser desenvolvidos analisando os propagadores menos influentes considerando diferentes estruturas topológicas e tipos de redes, métodos adaptativo mais elaborados e considerando outros processos de difusão.

CONCLUSÕES

Pesquisas relacionadas à propagação de informação compreendem diversos tópicos, tais como rumores, notícias, ideias, opiniões e a maneira que estes se espalham a grande escala com o passar do tempo. A difusão da informação é um processo ubíquo na sociedade, onde as pessoas estão interagindo e compartilhando informações o tempo todo, seja no transporte público, no trabalho, na escola, em casa ou na Internet (NOË; WHITAKER; ALLEN, 2016; TABACCHI *et al.*, 2017). Desta forma, a transmissão de informação em redes sociais e na sociedade é importante para cenários como marketing viral, campanhas políticas, educacionais ou ideológicas, entre outros.

Desde o surgimento da teoria de redes complexas (BARABÁSI; ALBERT, 1999; NEWMAN, 2010; CASTELLANO; FORTUNATO; LORETO, 2009), foram propostos modelos mais precisos para analisar processos de propagação ou de contágio social, tais como a formação de opiniões, tomada de decisões e propagação de rumores (CASTELLANO; FORTUNATO; LORETO, 2009; GUILLE *et al.*, 2013; PASTOR-SATORRAS *et al.*, 2015; KEMPE; KLEINBERG; TARDOS, 2015; PORTER; GLEESON, 2016; VEGA-OLIVEROS; COSTA; RODRIGUES, 2017b). Contudo, a abordagem geralmente adotada na literatura é uma representação simples onde todas as taxas ou probabilidades de propagação são idênticas para todos os vértices. Isto não é uma propriedade comumente observada em cenários reais, pois existem indivíduos que têm maior capacidade de convencer ou ser convencidos sobre alguma informação. Considerar taxas heterogêneas ou diferentes tipos de vértices na rede é um passo relevante na modelagem de propagação de informação. Ademais, outras questões permanecem em aberto, como o impacto de medidas topológicas, como a correlação de grau, na difusão de informação, identificação de vértices ou conjunto de vértices influentes, estratégias para aumentar ou reduzir a influência de um rumor ou informação, e muitas outras.

Neste trabalho de doutorado contribuímos na avaliação e quantificação de como a organização da rede influencia nos processos dinâmicos que são simulados sobre ela. Adotamos

várias medidas de redes complexas, modelos de redes e modelos de propagação. As simulações e análise dos modelos e métodos que aqui propomos englobaram diferentes tipos de redes, tanto artificiais como reais, em especial redes sociais, tecnológicas e de informação.

Fizemos avaliações em escala global, meso escala e escala local, visando estudar e diferenciar o comportamento da propagação em cada caso e identificar padrões na presença de propagadores mais ou menos influentes. Baseados na formulação teórica de modelos existentes, formalizamos e propomos um novo modelo heterogêneo de propagação de rumores, comparando os resultados teóricos com os resultados das simulações, bem como as abordagens clássicas. Analisamos a heterogeneidade das probabilidades de transmissão quando estão correlacionadas com medidas estruturais da rede, ou independentes da estrutura, com alguma distribuição de probabilidade como Poisson, lei de potência, entre outras. Como resultados, observamos a presença de propagadores mais influentes na dinâmica de rumor heterogênea, sendo estes os vértices com maior grau, centralidade de intermediação ou PageRank. Também, encontramos que a correlação da taxa de propagação com essas medidas produz os melhores resultados considerando a média de propagação da rede. Adicionalmente, mostramos que é possível, ao modificar as taxas de propagação de uma pequena fração dos propagadores mais influentes, controlar ou aumentar consideravelmente (entre 30% e 50%) o alcance da difusão da informação na rede.

Além disso, abordamos o problema de maximização de informação na rede, o qual busca encontrar o conjunto de η indivíduos ou sementes iniciais que maximizem a quantidade de pessoas “infectadas” na rede. Esse problema na literatura é considerado *NP*-difícil (KEMPE; KLEINBERG; TARDOS, 2003). Propusemos um método que seleciona os propagadores a partir da estrutura de comunidades da rede, com um custo computacional baixo ($\mathcal{O}(n \log(n))$) e que não depende do tamanho η . Mostramos que nossa proposta, além de ser performática e aplicável a problemas do mundo real, produz resultados estatisticamente semelhantes aos do método estado da arte.

Analisamos também o impacto que a correlação de grau-grau exerce na difusão de informação, considerando um conjunto de η propagadores iniciais. Observamos que as redes disassortativas ou com correlação negativa no grau, a melhor opção é selecionar os vértices mais centrais da rede. Ainda mais, existe um tamanho ótimo de η onde a seleção dos vértices com maior centralidade da rede produz um pico de infectados. Depois desse ponto, há sobreposição na influência dos propagadores e as densidades finais de informados começa a cair. Esse resultado indica que não necessariamente aumentar o número de propagadores iniciais aumentará a densidade final esperada de indivíduos informados.

Finalmente, contribuímos com a definição e identificação dos vértices menos influentes no processo de difusão de informação, tema que não tinha sido explorado na literatura. Encontramos que os vértices com menores valores de centralidade de autovetor são aqueles que menos propagam na rede. Como resultado não esperado, observamos que não necessariamente os

vértices com grau ou centralidade de intermediação baixa são os que menos propagam. Vértices com valores de medidas de centralidade acima da média da rede podem ser propagadores menos influentes, indicando que não se tratava de um problema trivial. Adicionalmente, mostramos a importância que os vértices menos influentes podem ter na propagação de informação. Mediante um método adaptativo aqui proposto, mostramos que ao selecionar uma pequena fração (0.5%) dos vértices menos influentes e reconectado apenas uma de suas arestas para um vértice central da rede, a capacidade de propagação desse e de todos os vértices da rede é aumentada. Isto poderia ser interpretado, como se os vértices menos influentes fossem “gargalos” na dinâmica.

8.1 Resultados e contribuições

8.1.1 Contribuições

O objetivo geral desta tese foi contribuir e complementar os estudos sobre dinâmicas de propagação de informação e rumores em redes complexas, considerando a topologia da rede. A seguir resumimos as principais contribuições deste trabalho de doutorado:

- Proposta de um modelo de propagação de rumores considerando probabilidades de transmissão e recuperação heterogêneas para os vértices da rede.
- Proposta de uma técnica de controle ou melhora na propagação considerando os vértices mais influentes da rede, podendo potencializar ou reduzir o nível de propagação.
- Contribuição com o entendimento de como a correlação de grau-grau na rede afeta a dinâmica e especificamente no problema de maximização de informação.
- Identificação, mediante abordagem topológica e orientada a aprendizado de máquina não supervisionado, dos vértices menos influentes da rede.
- Proposta de um método adaptativo para melhorar o potencial de propagação dos vértices menos influentes, sem alterar a distribuição de grau da rede. Mostramos que identificar os vértices menos influentes da rede é interessante, pois é possível aumentar a capacidade geral de propagação da rede.
- Avanço no entendimento sobre a relação entre a estrutura e a dinâmica, o que constitui um dos principais tópicos de redes complexas (BARRAT; BARTHÉLEMY; VESPIGNANI, 2008; NEWMAN, 2010).

8.1.2 Publicações

Publicações e submissões diretamente relacionados com o tema de pesquisa:

1. VEGA-OLIVEROS, D. A.; F COSTA, L.; RODRIGUES, F. A. Influence maximization in complex networks: the role of degree-degree correlation. Submetido em *Physical Review E*, 2017.
2. VEGA-OLIVEROS, D. A.; BERTON, L.; VAZQUEZ, F.; RODRIGUES, F. A. The Impact of Social Curiosity on Information Spreading on Networks. Aceito em *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2017.
3. VEGA-OLIVEROS, D. A.; F COSTA, L.; RODRIGUES, F. A. Rumor propagation with heterogeneous transmission in social networks. *Journal of Statistical Mechanics: Theory and Experiment*, v. 2017, n. 2, p. 023401, 2017.
4. VEGA-OLIVEROS, D. A.; BERTON, L. Spreader selection by community to maximize information diffusion in social networks. Em: *SIMBig'15*, 2015, páginas 73–82.
5. VEGA-OLIVEROS, D. A.; BERTON, L.; DE ANDRADE LOPES, A.; RODRIGUES, F. A. Influence maximization based on the least influential spreaders. Em: *Workshop on Social Influence Analysis. International Joint Conference on Artificial Intelligence (IJCAI 2015)*, 2015, páginas 3–8.

Publicações resultantes de colaborações do aluno em áreas co-relacionadas ao tema de pesquisa:

1. MÉNDEZ-BERMÚDEZ, J. A.; VEGA-OLIVEROS, D. A.; RODRIGUES, F. A. Multifractality in random networks with long-range spatial correlations. Submetido em *Physical Review E*, 2017.
2. BERTON, L.; VEGA-OLIVEROS, D. A. The Influence of the Degree in Graph-based Semi-Supervised Learning. Submetido em *Knowledge and Information Systems (KAIS)*, 2017.
3. BERTON, L.; VEGA-OLIVEROS, D. A.; VALVERDE-REBAZA, J.; SILVA, A. T.; LOPES, A. D. A. *Network sampling based on centrality measures for relational classification* Capítulo de livro: Springer International Publishing, páginas 43–56, 2017.
4. BERTON, L.; VEGA-OLIVEROS, D. A.; VALVERDE-REBAZA, J. C.; SILVA, A. T.; ANDRADE LOPES, A. Network sampling based on centrality measures for relational classification. Em: *SIMBig'16*, 2016, páginas 43–56.
5. VEGA-OLIVEROS, D. A.; BERTON, L.; EBERLE, A. M.; LOPES, A. D. A.; ZHAO, L. Regular graph construction for semi-supervised learning. *Journal of Physics: Conference Series*, v. 490, n. 1, p. 012022, 2014.

8.2 Limitações

A primeira limitação afrontada nesta proposta e evidenciada nos trabalhos da literatura é a dimensão temporal do processo. Embora se assuma um tempo discreto ou contínuo, no mundo real o efeito viral tem um tempo relativo e diferente em cada fase. Relacionado ao anterior também temos a dinâmica temporal da população. Por exemplo, a população cresce ou se movimenta de acordo com um processo de entrada e saída de indivíduos, assim, podem existir diferentes regimes de prevalência da informação dependendo da dinâmica desse público (BRITTON *et al.*, 2015). É possível que produtos ou ideias perdurem em um setor da população, mas o número de indivíduos de dado setor diminui lentamente com o tempo, fazendo com que a fração de potenciais clientes tenda a zero. Também, é possível que o público alvo seja mantido ou aumente, alcançando um equilíbrio onde a fração de potenciais seguidores convirja a um valor constante.

No caso da entrada e saída de indivíduos da população, a abordagem mais simples é considerar uma produção linear em que os novos indivíduos chegam sem contatos na rede (BRITTON *et al.*, 2015). Segundo alguma taxa, cada novo indivíduo adquire novos contatos fazendo ligações aleatórias. Porém, persiste a questão de como realmente fazer as conexões desses novos indivíduos. Por exemplo, no caso de novos adolescentes entrando no mercado, eles não seriam indivíduos isolados. Também, para o público de novos adultos saindo do grupo de adolescente, alguns deles podem perdurar por mais tempo e manter conexões com o grupo de adolescentes (exemplo, fans de revistas de quadrinhos).

Relacionado a população temos também uma limitação da probabilidade de interação dos indivíduos, a qual é independente de sua posição no sentido social ou espacial. Nas redes sociais reais existem estruturas locais nas quais observa-se uma alta transmissão de informação, como lugares de encontros, cinemas, etc., no caso de lugares físicos, e chats, blogs ou fóruns, no caso sites on-line. Ademais, não consideramos propriedades heterogêneas dos indivíduos que de fato afetam as probabilidades de contágio e a frequência das interações (NOË; WHITAKER; ALLEN, 2016; TABACCHI *et al.*, 2017), como faixa etária, sexo, condição socioeconômico, entre outras.

Nesse sentido, também existe uma forte limitação em relação à disponibilidade de dados reais que tenham características temporais e espaciais. Por exemplo, a maioria dos trabalhos empregam dados do Twitter, nos quais as entradas são frações de eventos de *Tweets* que difere do definido para uma rede social e tem muitas peculiaridades (LIU; CHEN, 2011). Desse modo, alguns dos resultados reportados são específicos, com métodos exclusivos para essa plataforma (SERRANO; IGLESIAS; GARIJO, 2015).

Outra limitação está na própria definição do que seria a maximização de informação e a definição do problema de maximização de informação. Segundo apresentado na literatura (RICHARDSON; DOMINGOS, 2002; KEMPE; KLEINBERG; TARDOS, 2003), a definição assume situações de propagação ideais (BONCHI, 2011). Como propor e analisar um problema

de maximização de influência que seja tratável e comparável a cenários reais? Também, e como foi discutido na Seção 3.1, os modelos de propagação fazem várias suposições sobre o modo como a informação se transmite, o padrão de conexões e a população, o comportamento dos indivíduos, seus estados discretos, entre outros. Contudo, não existem comparativos ou estudos entre os diferentes modelos de propagação, que estabeleçam um marco de referência para os métodos propostos e situações reais.

No ponto de vista mais técnico e estrutural, temos as limitações computacionais para as simulações e a representação dos dados. Mesmo assumindo o caso mais simples de redes não direcionadas e sem peso com uma população constante no tempo, o custo para manter em memória redes pequenas (com $N \leq 10^4$) e rodar os modelos e medidas de centralidade é alto. Assim, existe uma limitação nas técnicas e a capacidade computacional para poder fazer as simulações em redes de grande porte. Especificamente, para o problema de maximização de informação, a abordagem direta de calcular o conjunto de vértices que minimiza ou maximiza a difusão é um problema de explosão combinatória (*NP*-difícil).

A seguir resumimos os aspectos que poderiam ter sido considerados neste projeto e foram limitados:

- Tipos de modelos de propagação: adotamos a modelagem *SIR* para propagação de informação. Porém, outras modelagens considerando *SI*, *SIS*, *SEIS*, *SEIR*, *SEIRS*, entre outras, onde o estado *E* significa que os sujeitos estão informados, mas que ainda não começaram a compartilhar a informação, poderiam ser considerados.
- Tratamento da evolução do tempo: consideramos uma evolução temporal discreta, tanto nas simulações quanto na proposta do modelo, seguindo uma abordagem Markoviana onde o evento futuro só depende do estado do tempo atual e não do histórico de estados. Assim, os eventos acontecem segundo uma distribuição exponencial no tempo e poderia ser considerado ou proposto uma modelagem contínua.
- Tipo de tratamento da população: nas abordagens propostas consideramos uma população fixa, sem mudança no tamanho ou com flutuações ao redor do tamanho inicial com entrada e saída de indivíduos. Também, não foram considerados cliques ou classes de vértices, como diferença etária, sexo, classe socioeconômica, entre outras características na população.
- Flutuações no tempo: adotamos uma modelagem com probabilidades constantes durante a evolução temporal da dinâmica. Isto significa que todas as probabilidades foram mantidas iguais no tempo, sem considerar características sazonais, momentos de alta ou baixa nas condições de propagação sejam por fatores externos ou internos da dinâmica, entre outros.

8.3 Trabalhos Futuros

Além de propostas que lidem com as limitações levantadas na seção anterior, a seguir apresentamos algumas questões que podem ser pesquisadas futuramente.

8.3.1 *Relação entre a estrutura e a dinâmica*

Características da estrutura das redes *on-line* podem estar intrinsecamente ligadas com a personalidade dos seus usuários (TABACCHI *et al.*, 2017). Saber se a personalidade dos indivíduos coordena a estrutura da rede social ou vice-versa ainda não foi aclarado. Acredita-se que personalidade e a posição do indivíduo na rede social podem se reforçar entre si. Desse modo, analisar a relação entre os traços sociais e a estrutura da rede ajuda a melhorar o entendimento e aplicabilidade dos modelos de transmissão. A partir dessas análises podem ser exploradas corretamente mais propriedades reais nos modelos, como a apatia, perda de interesse ou disposição dos propagadores, memória ou esquecimento, não redundância na comunicação, difamação ou fofoca, curiosidade, entre muitas outras.

8.3.2 *Redes ponderadas*

Podem ser feitas análises em redes ponderadas e nos modelos de propagação, considerando: o impacto da distribuição dos pesos e das conexões; o papel na correlação dos pesos nas arestas, se são distribuídas aleatoriamente ou se os pesos estão correlacionados no nível local ou global; a relação entre o peso e as medidas de centralidade, por exemplo, nas redes sociais as pessoas com um número elevado de conexões têm pesos altos ou baixos nas suas conexões?; a relevância das conexões com baixo peso nas arestas, elas podem ser ignoradas ou elas são importantes para a difusão da informação em grupos locais densamente conectados. Adicional a isso, temos que os contatos ou pesos nas arestas podem mudar no tempo, pois a força dos contatos sociais não é nem contínuos nem permanentes.

8.3.3 *Dinâmica na população*

A estrutura da rede pode ser dinâmica, acompanhado as mudanças da população e conexões tais como, nascimentos, mortes, condições escolares, mudanças nas relações sociais, migração, mobilidade, mudanças sazonais ou econômicas de longo prazo, intervenções diretas como campanhas do governo, entre outras. Esse tipo de alterações afetam a estrutura da rede e a disseminação da informação por si. As alterações podem ser locais, como troca de conexões, duração dos contatos, entrada e saída equivalente de vértices ou globais, mudando propriedades estruturais da rede. Assim, cenários ou situações mais próximos a caso reais podem ser modelados, e esperam-se novas conclusões sobre a propagação e métodos mais precisos possam ser propostos.

REFERÊNCIAS

ADVOGATO network dataset – KONECT. 2016. Disponível em: <<http://konect.uni-koblenz.de/networks/advogato>>. Citado na página 141.

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. **Rev. Mod. Phys.**, American Physical Society, v. 74, n. 1, p. 47–97, jan 2002. Citado nas páginas 27, 44, 73 e 111.

AMARAL L.A.N, O. J. Complex networks - Augmenting the framework for the study of complex systems. **Eur. Phys. J. B**, v. 38, p. 147–162, MAR 2004. Citado na página 37.

ANDERSON, R. M.; MAY, R. M. **Infectious diseases of humans: Dynamics and control**. Oxford, UK: Oxford University Press, 1992. v. 108. 768 p. ISSN 02709139. Disponível em: <<http://doi.wiley.com/10.1002/hep.1840150131>>. Citado nas páginas 27, 30, 61, 62, 67 e 69.

APOLLONI, A.; GARGIULO, F. Diffusion Processes Through Social Groups' Dynamics. **Advances in Complex Systems**, World Scientific Publishing Company, v. 14, n. 02, p. 151–167, abr. 2011. ISSN 0219-5259. Disponível em: <<http://www.worldscientific.com/doi/abs/10.1142/S0219525911003037>>. Citado na página 27.

AXELROD, R. The Dissemination of Culture: A Model with Local Convergence and Global Polarization. **Journal of Conflict Resolution**, v. 41, n. 2, p. 203–226, abr. 1997. ISSN 0022-0027. Disponível em: <<http://jcr.sagepub.com/content/41/2/203>>. Citado na página 27.

BAE, J.; KIM, S. Identifying and ranking influential spreaders in complex networks by neighborhood coreness. **Physica A: Statistical Mechanics and its Applications**, v. 395, p. 549–559, fev. 2014. ISSN 03784371. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0378437113010406>>. Citado nas páginas 73, 74, 78 e 94.

BAILEY, N. T. **The Mathematical Theory of Infectious Diseases**. [S.l.]: Griffin, 1975. Citado na página 62.

BAJARDI, P.; POLETTO, C.; RAMASCO, J. J.; TIZZONI, M.; COLIZZA, V.; VESPIGNANI, A. Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic. **PloS one**, Public Library of Science, v. 6, n. 1, p. e16591, jan. 2011. ISSN 1932-6203. Disponível em: <<http://www.plosone.org/article/info:doi/10.1371/journal.pone.0016591>>. Citado na página 37.

BAKSHY, E.; ROSENN, I.; MARLOW, C.; ADAMIC, L. The role of social networks in information diffusion. In: **Proceedings of the 21st International Conference on World Wide Web**. New York, NY, USA: ACM, 2012. (WWW '12), p. 519–528. ISBN 978-1-4503-1229-5. Disponível em: <<http://doi.acm.org/10.1145/2187836.2187907>>. Citado nas páginas 27, 28, 94 e 105.

BARABÁSI, A.-L. The physics of the web. **Physics World**, p. 33–38, 2001. Citado nas páginas 27, 41 e 73.

- BARABÁSI, A.-L. The architecture of complexity: From network structure to human dynamics. **IEEE Control Systems Magazine**, v. 27, n. 4, p. 33–42, 2007. Citado nas páginas 28, 29, 46 e 73.
- _____. Scale-free networks: A decade and beyond. **Science**, v. 325, n. 5939, p. 412–413, 2009. Citado nas páginas 38, 44 e 78.
- BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **Science**, v. 286, p. 509–512, 1999. Citado nas páginas 19, 40, 44, 94, 99, 116, 122 e 149.
- BARRAT, A.; BARTHÉLEMY, M.; VESPIGNANI, A. **Dynamical processes on complex networks**. Cambridge University Press, 2008. Disponível em: <<http://www.cambridge.org/us/academic/subjects/physics/statistical-physics/dynamical-processes-complex-networks>>. Citado nas páginas 27, 28, 29, 30, 45, 55, 61, 63, 64, 65, 67, 69, 70, 71, 72, 75, 76, 83, 84, 87, 110, 140 e 151.
- BARRAT, A.; Weigt M. On the properties of small-world networks. **Eur. Phys. J. B** 13, p. 547–560, 2000. Citado na página 50.
- BARTHÉLEMY, M. Crossover from scale-free to spatial networks. **Europhysics Letters (EPL)**, IOP Publishing, v. 63, n. 6, p. 915–921, sep 2003. ISSN 0295-5075. Citado nas páginas 20, 45, 46, 99 e 100.
- BATAGELJ, V.; ZAVERSNIK, M. An O(m) algorithm for cores decomposition of networks. **Arxiv preprint cs/0310049**., 2003. Citado nas páginas 19 e 53.
- BERTOTTI, M.; BRUNNER, J.; MODANESE, G. The Bass diffusion model on networks with correlations and inhomogeneous advertising. **Chaos, Solitons & Fractals**, v. 90, p. 55–63, 2016. Citado na página 121.
- BETTENCOURT, L. M.; CINTRÓN-ARIAS, A.; KAISER, D. I.; CASTILLO-CHÁVEZ, C. The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. **Physica A: Statistical Mechanics and its Applications**, v. 364, p. 513–536, maio 2006. ISSN 03784371. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0378437105010113>>. Citado na página 75.
- BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D.-U. Complex Networks : Structure and Dynamics. **Phys. Rep.**, v. 424, n. 4-5, p. 175–308, 2006. Citado nas páginas 27, 28, 73, 81, 82, 91, 94 e 103.
- BOLKER, B.; GRENLELL, B. Chaos and biological complexity in measles dynamics. In: **Proc R Soc London Ser B**. [S.l.: s.n.], 1993. p. 251,75–81. Citado na página 61.
- BOLLOBÁS, B. Random graphs. **Cambridge Univ Pr.**, v. 73, 2001. Citado nas páginas 40 e 41.
- BONACICH, P. Factoring and weighting approaches to clique identification. **Journal of Mathematical Sociology**, v. 2, p. 113–120, 1972. Citado nas páginas 19 e 51.
- BONCHI, F. Influence propagation in social networks: A data mining perspective. In: **Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 02**. Washington, DC, USA: IEEE Computer Society, 2011. (WI-IAT '11), p. 2–. ISBN 978-0-7695-4513-4. Disponível em: <<http://dx.doi.org.ez67.periodicos.capes.gov.br/10.1109/WI-IAT.2011.292>>. Citado nas páginas 86 e 153.

- BORGE-HOLTHOEFER, J.; MELONI, S.; GONÇALVES, B.; MORENO, Y. Emergence of Influential Spreaders in Modified Rumor Models. **Journal of Statistical Physics**, v. 151, n. 1-2, p. 383–393, set. 2012. ISSN 0022-4715. Disponível em: <<http://link.springer.com/10.1007/s10955-012-0595-6>>. Citado nas páginas 80, 87, 89, 94 e 140.
- BORGE-HOLTHOEFER, J.; MORENO, Y. Absence of influential spreaders in rumor dynamics. **Physical Review E**, v. 85, n. 2, p. 026116, fev. 2012. ISSN 1539-3755. Disponível em: <<http://link.aps.org/doi/10.1103/PhysRevE.85.026116>>. Citado nas páginas 27, 30, 32, 75, 76, 77, 78, 86, 87, 121 e 124.
- BORGE-HOLTHOEFER, J.; RIVERO, A.; MORENO, Y. Locating privileged spreaders on an online social network. **Phys. Rev. E**, American Physical Society, v. 85, p. 066123, Jun 2012. Disponível em: <<http://link.aps.org/doi/10.1103/PhysRevE.85.066123>>. Citado na página 81.
- BORNHOLDT, S.; ROHLF, T. Topological evolution of dynamical networks: Global criticality from local dynamics. **Phys. Rev. Lett.**, American Physical Society, v. 84, p. 6114–6117, Jun 2000. Disponível em: <<http://link.aps.org/doi/10.1103/PhysRevLett.84.6114>>. Citado na página 145.
- BRANDES, U. A Faster Algorithm for Betweenness Centrality. **Journal of Mathematical Sociology**, v. 25, p. 163–177, 2001. Citado na página 39.
- BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. **Computer Networks and ISDN Systems**, V, p. 107–117, 1998. Citado nas páginas 19 e 52.
- BRITTON, T.; HOUSE, T.; LLOYD, A. L.; MOLLISON, D.; RILEY, S.; TRAPMAN, P. Five challenges for stochastic epidemic models involving global transmission. **Epidemics**, v. 10, p. 54–57, 2015. ISSN 1755-4365. Challenges in Modelling Infectious {Disease} Dynamics. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1755436514000280>>. Citado nas páginas 30, 31 e 153.
- BUONO, C.; VAZQUEZ, F.; MACRI, P. A.; BRAUNSTEIN, L. A. Slow epidemic extinction in populations with heterogeneous infection rates. **Physical Review E**, v. 88, n. 2, p. 022813, 2013. ISSN 1539-3755. Citado na página 86.
- BURT, R. S. **Structural holes: The social structure of competition**. Cambridge, MA: Harvard University Press, 1992. Citado nas páginas 19, 53 e 140.
- CAO, T.; WU, X.; WANG, S.; HU, X. Maximizing influence spread in modular social networks by optimal resource allocation. **Expert Systems with Applications**, v. 38, n. 10, p. 13128–13135, 2011. Citado na página 112.
- CASTELLANO, C. Social Influence and the Dynamics of Opinions: The Approach of Statistical Physics. **Managerial and Decision Economics**, v. 33, n. 5-6, p. 311–321, jul. 2012. ISSN 01436570. Disponível em: <<http://doi.wiley.com/10.1002/mde.2555>>. Citado na página 27.
- CASTELLANO, C.; FORTUNATO, S.; LORETO, V. Statistical Physics of Social Dynamics. **Reviews of Modern Physics**, v. 81, n. 2, p. 591–646, maio 2009. Citado nas páginas 27, 28, 29, 71, 77, 80, 85, 87, 94, 95, 110, 120, 140 e 149.
- CATALDI, M.; Di Caro, L.; SCHIFANELLA, C. Emerging topic detection on Twitter based on temporal and social terms evaluation. In: **Proceedings of the Tenth International Workshop on Multimedia Data Mining - MDMKDD '10**. New York, New York, USA: ACM Press, 2010. p. 1–10. ISBN 9781450302203. Citado na página 111.

- CHEN, D.-B.; GAO, H.; Lü, L.; ZHOU, T. Identifying influential nodes in large-scale directed networks: The role of clustering. **PLoS ONE**, Public Library of Science, v. 8, n. 10, p. e77455, 10 2013. Disponível em: <<http://dx.doi.org/10.1371/journal.pone.0077455>>. Citado nas páginas 29, 72 e 74.
- CHEN, W.; WANG, Y.; YANG, S. Efficient influence maximization in social networks. In: **Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09**. New York, New York, USA: ACM Press, 2009. p. 199. Citado na página 112.
- CHOWELL, G.; HYMAN, J. M.; EUBANK, S.; CASTILLO-CHAVEZ, C. Scaling laws for the movement of people between locations in a large city. **Phys. Rev. E**, v. 68, p. 066102, 2003. Citado na página 62.
- CIMINI, G.; MEDO, M.; ZHOU, T.; WEI, D.; ZHANG, Y.-C. Heterogeneity, quality, and reputation in an adaptive recommendation model. **Eur Phys JB**, v. 80, p. 201–8, 2011. Disponível em: <[http://refhub.elsevier.com/S1007-5704\(13\)00378-X/h0105](http://refhub.elsevier.com/S1007-5704(13)00378-X/h0105)>. Citado na página 79.
- CLAUSET, A.; NEWMAN, M. E. J.; MOORE, C. Finding community structure in very large networks. **Physical Review E**, v. 70, n. 3, p. 66111, 2004. Citado na página 58.
- CLAUSET, A.; SHALIZI, C. R.; NEWMAN, M. E. J. Power-law distributions in empirical data. **SIAM Review**, v. 51, n. 4, p. 661–703, 2009. Citado na página 44.
- COHEN, R.; EREZ, K.; BEN-AVRAHAM, D.; HAVLIN, S. Resilience of the Internet to Random Breakdowns. **Physical Review Letters**, American Physical Society, v. 85, n. 21, p. 4626+, November 2000. Citado na página 73.
- COLIZZA, V.; PASTOR-SATORRAS, R.; VESPIGNANI, A. Reaction-diffusion processes and metapopulation models in heterogeneous networks. **Nature Physics**, v. 3, n. 4, p. 276–282, mar 2007. ISSN 1745-2473. Disponível em: <<http://dx.doi.org/10.1038/nphys560>>. Citado na página 61.
- COSTA, L.; Oliveira Jr, O. N.; TRAVIESO, G.; RODRIGUES, F. A.; BOAS, P. R. V.; ANTIQUEIRA, L.; VIANA, M. P.; ROCHA, L. E. C. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. **Advances in Physics**, Taylor & Francis, v. 60, n. 3, p. 329–412, 2011. Citado nas páginas 27, 28 e 53.
- COSTA, L. D. F.; RODRIGUES, F. A.; TRAVIESO, G.; Villas Boas, P. R. Characterization of complex networks: A survey of measurements. **Advances in Physics**, v. 56, p. 167–242, 2007. Citado nas páginas 38, 41, 46, 47, 56 e 64.
- DALEY, D. J.; KENDALL, D. G. Epidemics and Rumours. **Nature**, v. 204, n. 4963, p. 1118–1118, dez. 1964. ISSN 0028-0836. Disponível em: <<http://dx.doi.org/10.1038/2041118a0>>. Citado nas páginas 19, 75, 76, 77, 82, 94 e 121.
- DANO, L.; FORD, A. P.; HOUSE, T.; JEWELL, C. P.; KEELING, M. J.; ROBERTS, G. O.; ROSS, J. V.; VERNON, M. C. Networks and the Epidemiology of Infectious Disease. **Interdisciplinary Perspectives on Infectious Diseases**, 2011. Citado na página 62.
- DEMSAR, J. Statistical comparisons of classifiers over multiple data sets. **J. Mach. Learn. Res.**, JMLR.org, v. 7, p. 1–30, dez. 2006. ISSN 1532-4435. Disponível em: <<http://dl.acm.org/citation.cfm?id=1248547.1248548>>. Citado nas páginas 135 e 136.

DIEKMANN, O.; HEESTERBEEK, J. A. P. **Mathematical Epidemiology of Infectious**. [S.l.]: Wiley, 2000. Citado na página 62.

DINH, T.; NGUYEN, H.; GHOSH, P.; MAYO, M. Computational Social Networks: 4th International Conference, CSoNet 2015, Beijing, China, August 4-6, 2015, Proceedings. In: _____. Cham: Springer International Publishing, 2015. cap. Social Inf, p. 84–103. ISBN 978-3-319-21786-4. Citado nas páginas 125, 126 e 136.

DODDS, P.; WATTS, D. Universal behavior in a generalized model of contagion. **Phys Rev Lett**, v. 92, p. 218701, 2004. Citado nas páginas 79 e 80.

DRAIEF, M.; MASSOULI, L. **Epidemics and rumours in complex networks**. [S.l.]: Cambridge University Press, 2010. Citado na página 30.

ERDÖS, P.; RÉNYI, A. On random graphs. **Publicationes Mathematicae**, v. 6, n. 1, p. 290–297, 1959. Citado nas páginas 19, 40 e 98.

EUBANK, S.; GUCLU, H.; KUMAR, V. S. A.; SRINIVASAN, A.; TOROCZKAI, Z.; WANG, N. Modelling disease outbreaks in realistic urban social networksNo Title. **Nature**, v. 429, p. 180–184., 2004. Citado na página 62.

FORTUNATO, S.; HRIC, D. Community detection in networks: A user guide. **Physics Reports**, Elsevier, v. 659, p. 1–44, 2016. Citado na página 113.

FREEMAN, L. C. A set of measures of centrality based on betweenness. **Sociometry**, v. 40, p. 35–41, 1977. Citado nas páginas 19 e 48.

_____. Centrality in Social Networks: Conceptual Clarification. **Social Networks**, v. 1, p. 215–239, 1979. Citado na página 73.

FUNK, S.; SALATHÉ, M.; JANSEN, V. A. A. Modelling the influence of human behaviour on the spread of infectious diseases: a review. **Journal of the Royal Society, Interface / the Royal Society**, v. 7, n. 50, p. 1247–56, set. 2010. ISSN 1742-5662. Disponível em: <<http://rsif.royalsocietypublishing.org/content/early/2010/05/25/rsif.2010.0142.full.html>>. Citado nas páginas 28 e 94.

GALSTYAN, A.; MUSOYAN, V.; COHEN, P. Maximizing influence propagation in networks with community structure. **Physical review. E, Statistical, nonlinear, and soft matter physics**, American Physical Society, v. 79, n. 5 Pt 2, p. 056102, may 2009. ISSN 1539-3755. Disponível em: <<http://journals.aps.org/pre/abstract/10.1103/PhysRevE.79.056102>>. Citado nas páginas 112 e 125.

GANI, J. The Maki-Thompson rumour model: a detailed analysis. **Environmental Modelling & Software**, v. 15, n. 8, p. 721–725, 2000. Citado na página 76.

GIRVAN, M.; NEWMAN, M. E. J. Community structure in social and biological networks. In: **Proc. Nat. Acad. Sci. USA** 99. [S.l.: s.n.], 2002. p. 7821–7826. Citado nas páginas 27 e 57.

GONZÁLEZ-BAILÓN, S.; Borge-Holthoefer, Javier Rivero; MORENO, Y. The dynamics of Protest Recruitment through an Online Network. **Scientific Reports**, v. 1:197, p. 1–7, 2011. Citado nas páginas 28, 94 e 140.

GOOGLE+ network dataset – KONECT. 2016. Disponível em: <<http://konect.uni-koblenz.de/networks/ego-gplus>>. Citado na página 141.

- GRENFELL, B. (Meta)population dynamics of infectious diseases. **Trends in Ecology & Evolution**, v. 12, n. 10, p. 395–399, out. 1997. ISSN 01695347. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0169534797011749>. Citado nas páginas 28 e 61.
- GROSS, T.; BLASIU, B. Adaptive coevolutionary networks: a review. **Journal of the Royal Society, Interface / the Royal Society**, v. 5, n. 20, p. 259–71, mar. 2008. ISSN 1742-5689. Disponível em: <http://rsif.royalsocietypublishing.org/content/5/20/259.short>. Citado na página 145.
- GROSS, T.; CARLOS, D.; BERND, B. Epidemic Dynamics on an Adaptive Network. **Physical Review Letters**, v. 96, n. 20, p. 208701–4., 2006. Citado na página 145.
- GU, J.; LI, W.; CAI, X. The effect of the forget-remember mechanism on spreading. **Eur. Phys. JB**, v. 62, p. 247–255, 2008. Citado na página 80.
- GUERRA, B.; nES, J. G.-G.; LATROA, V. Adaptive Growing Networks Coevolving with the Spread of Diseases. **International Journal of Bifurcation and Chaos**, World Scientific Publishing Company, v. 22, n. 07, p. 1250168, jul. 2012. ISSN 0218-1274. Disponível em: <http://www.worldscientific.com/doi/abs/10.1142/S0218127412501684>. Citado na página 145.
- GUILLE, A.; HACID, H.; FAVRE, C.; ZIGHED, D. A. Information diffusion in online social networks: A survey. **ACM SIGMOD Record**, ACM, v. 42, n. 1, p. 17, 5 2013. ISSN 01635808. Citado nas páginas 85, 86 e 149.
- GUIMERA, R.; DANON, L.; DIAZ-GUILERA, A.; GIRALT, F.; ARENAS, A. Self-similar community structure in a network of human interactions. **Physical Review E**, v. 68, p. 2003, 2003. Citado nas páginas 91 e 113.
- HA, J.; KIM, S.-W.; FALOUTSOS, C.; PARK, S. An Analysis on Information Diffusion through BlogCast in a Blogosphere. **Information Sciences**, v. 290, p. 45–62, jan 2015. ISSN 00200255. Citado nas páginas 94 e 103.
- HAN, Q.; WEN, H.; WU, J.; REN, M. Computational social networks: 4th international conference, csonet 2015, beijing, china, august 4-6, 2015, proceedings. In: _____. Cham: Springer International Publishing, 2015. cap. Rumor Spreading and Security Monitoring in Complex Networks, p. 48–59. ISBN 978-3-319-21786-4. Disponível em: http://dx.doi.org/10.1007/978-3-319-21786-4_5. Citado nas páginas 86 e 87.
- HÉBERT-DUFRESNE, L.; ALLARD, A.; YOUNG, J.-G.; DUBÉ, L. J. Global efficiency of local immunization on complex networks. **Scientific reports**, Nature Publishing Group, v. 3, p. 2171, jan. 2013. ISSN 2045-2322. Disponível em: <http://www.nature.com/srep/2013/130710/srep02171/full/srep02171.html>. Citado nas páginas 73, 74, 110, 111, 113 e 125.
- HETHCOTE, H. W. The Mathematics of Infectious Diseases. **SIAM REVIEW**, Society for Industrial and Applied Mathematics, ago. 2006. Disponível em: <http://epubs.siam.org/doi/abs/10.1137/S0036144500371907>. Citado nas páginas 37, 60, 61, 71 e 94.
- HOSSEINI-POZVEH, M.; ZAMANIFAR, K.; NAGHSH-NILCHI, A. R. A community-based approach to identify the most influential nodes in social networks. **Journal of Information Science**, p. 0165551515621005–, feb 2016. ISSN 0165-5515. Disponível em: <http://jis.sagepub.com/content/early/2016/02/04/0165551515621005.abstract>. Citado na página 121.

HUANG, S.; LV, T.; ZHANG, X.; YANG, Y.; ZHENG, W.; WEN, C. Identifying node role in social network based on multiple indicators. **PloS one**, Public Library of Science, v. 9, n. 8, p. e103733, jan 2014. ISSN 1932-6203. Disponível em: <<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0103733><http://dx.plos.org/10.1371/journal.pone.0103733>>. Citado nas páginas 111, 120 e 125.

KADUSHIN, C. Who benefits from network analysis: ethics of social network research. **Social Networks**, v. 27, n. 2, p. 139 – 153, 2005. ISSN 0378-8733. <ce:title>Ethical Dilemmas DILEMMAS in Social Network Research</ce:title>. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S037887330500016X>>. Citado na página 37.

KANDHWAY, K.; KURI, J. Using Node Centrality and Optimal Control to Maximize Information Diffusion in Social Networks. **IEEE Transactions on Systems, Man, and Cybernetics: Systems**, PP, n. 99, p. 1–12, 2016. ISSN 2168-2216. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7432048>>. Citado nas páginas 111, 120 e 125.

KEELING, M. J.; EAMES, K. T. D. Networks and epidemic models. **Journal of the Royal Society Interface**, The Royal Society, v. 2, n. 4, p. 295–307, 2005. Citado nas páginas 27, 29, 30, 61, 62 e 71.

KEELING, M. J.; ROHANI, P. **Modeling infectious diseases in humans and animals**. [S.l.]: Princeton University Press, 2008. Citado nas páginas 61, 94 e 95.

KEMPE, D.; KLEINBERG, J.; TARDOS Éva. Maximizing the spread of influence through a social network. **Theory of Computing**, Theory of Computing, v. 11, n. 4, p. 105–147, 2015. Disponível em: <<http://www.theoryofcomputing.org/articles/v011a004>>. Citado nas páginas 31, 85, 111, 112, 120, 121, 125, 126, 136 e 149.

KEMPE, D.; KLEINBERG, J. M.; TARDOS, É. Maximizing the Spread of Influence Through a Social Network. In: GETOOR, L.; SENATOR, T. E.; DOMINGOS, P.; FALOUTSOS, C. (Ed.). **KDD**. [S.l.]: ACM, 2003. p. 137–146. ISBN 1-58113-737-0. Citado nas páginas 19, 28, 29, 37, 73, 85, 86, 110, 111, 120, 121, 126, 150 e 153.

KERMACK, W. O.; MCKENDRICK, A. A Contribution to the Mathematical Theory of Epidemics. **Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character**, v. 115, n. 772, p. 700–721, 1927. Citado na página 62.

KISS, I. Z.; GREEN, D. M.; KAO, R. R. The effect of network mixing patterns on epidemic dynamics and the efficacy of disease contact tracing. **Journal of The Royal Society Interface**, The Royal Society, v. 5, n. 24, p. 791–799, 2008. Citado na página 120.

KITSAK, M.; GALLOS, L. K.; HAVLIN, S.; LILJEROS, F.; MUCHNIK, L.; STANLEY, H. E.; MAKSE, H. A. Identification of influential spreaders in complex networks. **Nature Physics**, Nature Publishing Group, v. 6, n. 11, p. 888–893, ago. 2010. ISSN 1745-2473. Disponível em: <http://www.nature.com/nphys/journal/v6/n11/fig_tab/nphys1746_ft.html>. Citado nas páginas 27, 29, 30, 32, 33, 53, 72, 73, 74, 77, 78, 94, 110, 111, 120, 125 e 140.

KOOPMAN, J. Controlling smallpox. **Science**, v. 298, n. 5597, p. 1342–1344, 2002. Disponível em: <<http://www.sciencemag.org/content/298/5597/1342.2.short>>. Citado na página 28.

- KOOPMAN, J. S. Infection transmission science and models. **Japanese journal of infectious diseases**, v. 58, n. 6, p. S3–8, dez. 2005. ISSN 1344-6304. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/16377860>>. Citado nas páginas 27 e 60.
- KREBS, V. Mapping Networks of Terrorist Cells. **Connections**, v. 24, n. 3, p. 43–52, 2002. Citado na página 37.
- KUNEGIS, J. **Hamsterster full network dataset – KONECT**. 2016. Disponível em: <<http://konect.uni-koblenz.de/networks/petster-hamster>>. Citado na página 141.
- _____. KONECT – The Koblenz Network Collection. In: **Proc. Int. Web Observatory Workshop**. [s.n.], 2017. p. 1343–1350. Disponível em: <<http://userpages.uni-koblenz.de/~kunegis/paper/kunegis-koblenz-network-collection.pdf>>. Citado nas páginas 90, 91 e 113.
- KUPERMAN, M. Cultural propagation on social networks. **Physical Review E**, v. 73, n. 4, p. 046139, abr. 2006. ISSN 1539-3755. Disponível em: <<http://link.aps.org/doi/10.1103/PhysRevE.73.046139>>. Citado nas páginas 27 e 28.
- LATORA, V.; MARCHIORI, M. Efficient behavior of small-world networks. **Phys. Rev. Lett.**, v. 87, n. 19, p. 198701, 2001. Citado na página 56.
- LAWYER, G. Measuring node spreading power by expected cluster degree. p. 4, set. 2012. Disponível em: <<http://arxiv.org/abs/1209.6600>>. Citado nas páginas 74 e 110.
- LESKOVEC, J.; KLEINBERG, J.; FALOUTSOS, C. Graphs over time: Densification laws, shrinking diameters and possible explanations. In: **Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining**. New York, NY, USA: ACM, 2005. (KDD '05), p. 177–187. ISBN 1-59593-135-X. Disponível em: <<http://doi.acm.org/10.1145/1081870.1081893>>. Citado na página 91.
- LI, W.; TANG, S.; PEI, S.; YAN, S.; JIANG, S.; TENG, X.; ZHENG, Z. The rumor diffusion process with emerging independent spreaders in complex networks. **Physica A: Statistical Mechanics and its Applications**, v. 397, p. 121–128, mar. 2014. ISSN 03784371. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0378437113010637>>. Citado nas páginas 32, 77 e 79.
- LIU, C.; ZHANG, Z.-K. Information spreading on dynamic social networks. **Communications in Nonlinear Science and Numerical Simulation**, v. 19, n. 4, p. 896–904, abr. 2014. ISSN 10075704. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S100757041300378X>>. Citado nas páginas 79 e 146.
- LIU, D.; CHEN, X. Rumor propagation in online social networks like twitter – a simulation study. In: **2011 Third International Conference on Multimedia Information Networking and Security**. [S.l.: s.n.], 2011. p. 278–282. ISSN 2162-8998. Citado na página 153.
- LOGLISCI, C.; CECI, M.; MALERBA, D. Relational mining for discovering changes in evolving networks. **Neurocomputing**, v. 150, p. 265–288, feb 2015. ISSN 09252312. Citado na página 95.
- LOPES, H.; PINTO, H. S.; FRANCISCO, A. P. Complex networks vii: Proceedings of the 7th workshop on complex networks complenet 2016. In: _____. Cham: Springer International Publishing, 2016. cap. Influence and Sentiment Homophily on Twitter Social Circles, p. 349–361. ISBN 978-3-319-30569-1. Disponível em: <http://dx.doi.org/10.1007/978-3-319-30569-1_27>. Citado nas páginas 19, 110 e 112.

LUSSEAU, D.; SCHNEIDER, K.; BOISSEAU, O.; HAASE, P.; SLOOTEN, E.; DAWSON, S. M. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. **Behavioral Ecology and Sociobiology**, v. 54, p. 396–405, 2003. Citado nas páginas 13 e 49.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: **In 5-th Berkeley Symposium on Mathematical Statistics and Probability**. [S.l.: s.n.], 1967. p. 281–297. Citado na página 144.

MAKI, D. P.; THOMPSON, M. **Mathematical Models and Applications, with Emphasis on the Social, Life, and Management Sciences**. [S.l.]: Prentice-Hall, 1973. Citado nas páginas 15, 19, 75, 76, 77, 82, 86, 87, 88, 94, 115, 121 e 137.

MASSA, P.; SALVETTI, M.; TOMASONI, D. Bowling alone and trust decline in social network sites. In: **Proc. Int. Conf. Dependable, Automatic and Secure Computing**. [S.l.: s.n.], 2009. p. 658–663. Citado nas páginas 91 e 113.

MCAULEY, J.; LESKOVEC, J. Learning to discover social circles in ego networks. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2012. p. 548–556. Citado na página 90.

MELONI, S.; ARENAS, A.; GÓMEZ, S.; BORGE-HOLTHOEFER, J.; MORENO, Y. Modeling epidemic spreading in complex networks: Concurrency and traffic. In: THAI, M. T.; PARDALOS, P. M. (Ed.). **Handbook of Optimization in Complex Networks**. Springer US, 2012, (Springer Optimization and Its Applications). p. 435–462. ISBN 978-1-4614-0753-9. Disponível em: <http://dx.doi.org/10.1007/978-1-4614-0754-6_15>. Citado nas páginas 72, 74, 76 e 89.

MELONI, S.; ARENAS, A.; MORENO, Y. Traffic-driven Epidemic Spreading in Finite-size Scale-free Networks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 106, n. 40, p. 16897–902, oct 2009. Citado na página 95.

MILGRAM, S. The Small World Problem. **Psychology Today**, v. 1, n. 1, p. 61–67, 1967. Citado nas páginas 42 e 90.

MILLER, J. C.; KISS, I. Z. Epidemic spread in networks: Existing methods and current challenges. mar. 2014. Disponível em: <<http://arxiv.org/abs/1403.1957>>. Citado nas páginas 29 e 71.

MORENO, Y.; NEKOVEE, M.; PACHECO, A. F. Dynamics of rumor spreading in complex networks. **Physical Review E**, AMER PHYSICAL SOC, ONE PHYSICS ELLIPSE, COLLEGE PK, MD 20740-3844 USA, v. 69, n. 6, p. 066130, jun 2004. ISSN 1539-3755. Citado nas páginas 27, 32, 77, 78, 87, 94, 110 e 140.

MORENO, Y.; NEKOVEE, M.; VESPIGNANI, A. Efficiency and reliability of epidemic data dissemination in complex networks. **Physical Review E**, v. 69, n. 5, p. 055101, maio 2004. ISSN 1539-3755. Disponível em: <<http://link.aps.org/doi/10.1103/PhysRevE.69.055101>>. Citado nas páginas 77, 78 e 121.

MORENO, Y.; PASTOR-SATORRAS, R.; VESPIGNANI, A. Epidemic outbreaks in complex heterogeneous networks. **The European Physical Journal B**, v. 26, n. 4, p. 521–529, abr. 2002. ISSN 1434-6028. Disponível em: <<http://arxiv.org/abs/cond-mat/0107267http://www.springerlink.com/index/10.1140/epjb/e20020122>>. Citado na página 70.

MORONE, F.; MAKSE, H. A. Influence maximization in complex networks through optimal percolation. **Nature**, Nature Publishing Group, 2015. Citado na página 120.

NÁ, M. B.; PASTOR-SATORRAS, R.; DÍAZ-GUILERA, A.; ARENAS, A. Models of social networks based on social distance attachment. **Phys. Rev. E**, American Physical Society, v. 70, p. 056122, Nov 2004. Disponível em: <<http://link.aps.org/doi/10.1103/PhysRevE.70.056122>>. Citado na página 92.

ná, M. B.; PASTOR-SATORRAS, R.; VESPIGNANI, A. Absence of Epidemic Threshold in Scale-Free Networks with Degree Correlations. **Physical Review Letters**, v. 90, n. 2, p. 028701, jan. 2003. ISSN 0031-9007. Disponível em: <<http://link.aps.org/doi/10.1103/PhysRevLett.90.028701>>. Citado nas páginas 30, 70, 72, 120 e 121.

NEKOVÉE, 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, jan. 2007. ISSN 03784371. Disponível em: <<http://arxiv.org/abs/0807.1458>>. Citado nas páginas 27, 30, 72, 75, 77, 78, 80 e 121.

NEWMAN, M. Assortative Mixing in Networks. **Physical Review Letters**, v. 89, n. 20, p. 208701, out. 2002. ISSN 0031-9007. Disponível em: <<http://link.aps.org/doi/10.1103/PhysRevLett.89.208701>>. Citado nas páginas 30, 72 e 78.

_____. Spread of epidemic disease on networks. **Physical Review E**, v. 66, n. 1, p. 016128, jul. 2002. ISSN 1063-651X. Disponível em: <<http://link.aps.org/doi/10.1103/PhysRevE.66.016128>>. Citado nas páginas 28, 61, 62 e 70.

_____. **Networks: an introduction**. [S.l.]: Oxford University Press, Inc., 2010. Citado nas páginas 27, 41, 62, 101, 149 e 151.

NEWMAN, M. E. J. The structure of scientific collaboration networks. In: **Natl. Acad. Sci. USA**. [S.l.: s.n.], 2001. p. 404 – 409. Citado nas páginas 91 e 113.

_____. The structure and function of complex networks. **SIAM Review**, v. 45, p. 1–8, 2003. Citado nas páginas 27, 38, 40, 41, 44, 46, 54, 73, 89, 90, 94 e 103.

_____. Fast algorithm for detecting community structure in networks. **Physical Review E**, v. 69, n. 3, p. 66133, 2004. Citado nas páginas 57, 58, 59 e 113.

_____. Finding community structure in networks using the eigenvectors of matrices. **Physical Review E**, v. 74, n. 3, p. 36104, 2006. Citado nas páginas 27, 57 e 58.

_____. **Personal Website Network Data**: <http://www-personal.umich.edu/~mejn/netdata/>. 2013. Citado na página 91.

NOË, N.; WHITAKER, R. M.; ALLEN, S. M. Personality homophily and the local network characteristics of facebook. In: **2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)**. [S.l.: s.n.], 2016. p. 386–393. Citado nas páginas 29, 149 e 153.

PACHECO, J.; TRAULSEN, A.; NOWAK, M. Coevolution of strategy and structure in complex networks with dynamical linking. **Phys Rev Lett**, v. 97(25), p. 258103, 2006. Disponível em: <[http://refhub.elsevier.com/S1007-5704\(13\)00378-X/h0140](http://refhub.elsevier.com/S1007-5704(13)00378-X/h0140)>. Citado nas páginas 79 e 146.

PASTOR-SATORRAS, R.; CASTELLANO, C.; Van Mieghem, P.; VESPIGNANI, A. Epidemic processes in complex networks. **Reviews of Modern Physics**, v. 87, n. 3, p. 925–979, aug 2015. ISSN 0034-6861. Disponível em: <<http://link.aps.org/doi/10.1103/RevModPhys.87.925>>. Citado nas páginas 30, 31, 86, 87, 94, 95, 98, 100, 120, 121 e 149.

PASTOR-SATORRAS, R.; VESPIGNANI, A. Epidemic Spreading in Scale-Free Networks. **Physical Review Letters**, v. 86, n. 14, p. 3200–3203, abr. 2001. ISSN 0031-9007. Disponível em: <<http://link.aps.org/doi/10.1103/PhysRevLett.86.3200>>. Citado nas páginas 30, 68, 71, 72, 73, 94 e 111.

PEI, S.; MAKSE, H. A. Spreading dynamics in complex networks. **Journal of Statistical Mechanics: Theory and Experiment**, IOP Publishing, v. 2013, n. 12, p. P12002, dez. 2013. ISSN 1742-5468. Disponível em: <<http://iopscience.iop.org/1742-5468/2013/12/P12002/article/>>. Citado nas páginas 72, 73, 74, 78, 110, 111, 120 e 125.

PELLIS, L.; BALL, F.; BANSAL, S.; EAMES, K.; HOUSE, T.; ISHAM, V.; TRAPMAN, P. Eight challenges for network epidemic models. **Epidemics**, v. 10, p. 58–62, 2015. ISSN 1755-4365. Challenges in Modelling Infectious {Disease} Dynamics. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1755436514000334>>. Citado nas páginas 30 e 31.

PORTER, M. A.; GLEESON, J. P. **Dynamical Systems on Networks: A Tutorial**. 1. ed. [S.l.]: Springer, 2016. 79 p. (Frontiers in Applied Dynamical Systems: Reviews and Tutorials). ISBN 9783319266404. Citado nas páginas 85, 86 e 149.

READ, J. M.; KEELING, M. J. Disease evolution on networks: the role of contact structure. **Proceedings. Biological sciences / The Royal Society**, v. 270, n. 1516, p. 699–708, abr. 2003. ISSN 0962-8452. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1691304&tool=pmcentrez&rendertype=abstract>>. Citado nas páginas 61 e 145.

RICHARDSON, M.; DOMINGOS, P. Mining knowledge-sharing sites for viral marketing. In: **Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02**. New York, New York, USA: ACM Press, 2002. p. 61. ISBN 158113567X. Disponível em: <<http://dl.acm.org/citation.cfm?id=775047.775057>>. Citado nas páginas 28, 29, 73, 110, 111, 121 e 153.

RUSSELL, B. **Marriage and Morals**. Liveright, 1970. (A Bantam classic). ISBN 9780871402110. Disponível em: <<https://books.google.com.br/books?id=g4NgDwJ4adcC>>. Citado na página 7.

RVACHEV, L.; LONGINI, I. A mathematical model for the global spread of influenza. **Math Biosci**, v. 75, p. 3 – 22, 1985. Citado na página 61.

SABIDUSSI, G. The centrality index of a graph. **Psychometrika**, v. 31, p. 581–603, 1996. Citado nas páginas 19, 50 e 140.

SANTOS, F.; PACHECO, J.; LENAERTS, T. Cooperation prevails when individuals adjust their social ties. **PLoS Comput Biol**, v. 2(10), n. 10, p. e140, 2006. Disponível em: <[http://refhub.elsevier.com/S1007-5704\(13\)00378-X/h0145](http://refhub.elsevier.com/S1007-5704(13)00378-X/h0145)>. Citado nas páginas 79 e 146.

SAYAMA, H.; PESTOV, I.; SCHMIDT, J.; BUSH, B. J.; WONG, C.; YAMANOI, J.; GROSS, T. Modeling complex systems with adaptive networks. **Computers & Mathematics with Applications**, v. 65, n. 10, p. 1645–1664, maio 2013. ISSN 08981221. Disponível em:

<http://www.sciencedirect.com/science/article/pii/S0898122112007018>>. Citado na página 145.

SEIDMAN, S. Network structure and minimum degree. **Social networks**, v. 5, n. 3, p. 269–287, 1983. Citado na página 52.

SERRANO, E.; IGLESIAS, C. A.; GARIJO, M. A survey of twitter rumor spreading simulations. In: _____. **Computational Collective Intelligence: 7th International Conference, ICCCI 2015, Madrid, Spain, September 21-23, 2015, Proceedings, Part I**. Cham: Springer International Publishing, 2015. p. 113–122. ISBN 978-3-319-24069-5. Disponível em: http://dx.doi.org/10.1007/978-3-319-24069-5_11>. Citado na página 153.

SIKIC, M.; LANCIC, A.; ANTULOV-FANTULIN, N.; STEFANCIC, H. Epidemic centrality- is there an underestimated epidemic impact of network peripheral nodes? **The European Physical Journal B**, Springer Berlin Heidelberg, v. 86, n. 10, p. 1–13, 2013. ISSN 1434-6028. Disponível em: <http://dx.doi.org/10.1140/epjb/e2013-31025-5>>. Citado na página 74.

STIEGLITZ, S.; DANG-XUAN, L. Emotions and information diffusion in social media - sentiment of microblogs and sharing behavior. **J. of Management Information Systems**, v. 29, n. 4, p. 217–248, 2013. Disponível em: <http://www.jmis-web.org/articles/904>>. Citado nas páginas 86 e 110.

TABACCHI, M. E.; CACI, B.; CARDACI, M.; PERTICONE, V. Early usage of pokémon go and its personality correlates. **Computers in Human Behavior**, v. 72, p. 163 – 169, 2017. ISSN 0747-5632. Citado nas páginas 28, 29, 86, 149, 153 e 155.

TEDXZURICH. **Pokémon GO is now the biggest mobile game in U.S. history**. 2017. Disponível em: <http://tedxzurich.com>>. Citado na página 28.

VEGA-OLIVEROS, D. A.; BERTON, L. Spreader selection by community to maximize information diffusion in social networks. In: VENTURA, J. A. L.; ALATRISTA-SALAS, H. (Ed.). **SIMBig'15**. CEUR-WS.org, 2015. (CEUR Workshop Proceedings, v. 1478), p. 73–82. Disponível em: <http://dblp.uni-trier.de/db/conf/simbig/simbig2015.html#Vega-OliverosB15>>. Citado nas páginas 33, 34, 102, 109, 115, 117 e 125.

VEGA-OLIVEROS, D. A.; BERTON, L.; LOPES, A. D. A.; RODRIGUES, F. A. Influence maximization based on the least influential spreaders. In: **Workshop on Social Influence Analysis. International Joint Conference on Artificial Intelligence (IJCAI 2015)**. Aachen, Germany, Germany: CEUR-WS.org, 2015. (SocInf'15), p. 3–8. Disponível em: <http://dl.acm.org/citation.cfm?id=2907168.2907171>>. Citado nas páginas 33, 35, 142, 143 e 147.

VEGA-OLIVEROS, D. A.; COSTA, L. da F.; RODRIGUES, F. A. Influence maximization in complex networks: the role of degree-degree correlation. **Arxiv preprint cs/1846515**., 2017. Citado nas páginas 33, 34, 35, 120, 123, 126, 127, 129, 130, 131, 132 e 135.

_____. Rumor propagation with heterogeneous transmission in social networks. **Journal of Statistical Mechanics: Theory and Experiment**, v. 2017, n. 2, p. 023401, 2017. Disponível em: <http://stacks.iop.org/1742-5468/2017/i=2/a=023401>>. Citado nas páginas 33, 34, 86, 93, 98, 99, 102, 103, 104, 107, 111, 120 e 149.

VESPIGNANI, A. Modelling dynamical processes in complex socio-technical systems. **Nature Physics**, Nature Publishing Group, v. 8, n. 1, p. 32–39, dez. 2011. ISSN 1745-2473. Disponível em: <<http://dx.doi.org/10.1038/nphys2160>>. Citado nas páginas 27, 29, 62, 72 e 77.

VIEIRA, I. T.; SENNA, V. de; PEREIRA, H. B. d. B. A small world model for the spread of HIV infection. **Pesquisa Operacional**, v. 31, n. 2, p. 373–389, ago. 2011. ISSN 0101-7438. Disponível em: <<http://www.scopus.com/inward/record.url?eid=2-s2.0-80051779155&partnerID=tZOtx3y1http://www.scielo.br/pdf/pope/v31n2/a09v31n2.pdf>>. Citado na página 28.

VINH, N. X.; EPPS, J.; BAILEY, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. **J. Mach. Learn. Res.**, JMLR.org, v. 11, p. 2837–2854, dez. 2010. ISSN 1532-4435. Disponível em: <<http://dl.acm.org/citation.cfm?id=1756006.1953024>>. Citado na página 132.

VISWANATH, B.; MISLOVE, A.; CHA, M.; GUMMADI, K. P. On the evolution of user interaction in Facebook. In: **Proc. Workshop on Online Social Networks**. [S.l.: s.n.], 2009. p. 37–42. Citado na página 29.

WANG, B.; TANG, H.; GUO, C.; XIU, Z. Entropy optimization of scale-free networks' robustness to random failures. **Physica A: Statistical Mechanics and its Applications**, v. 363, n. 2, p. 591–596, 2006. Citado na página 55.

WANG, Y.; CONG, G.; SONG, G.; XIE, K. Community-based greedy algorithm for mining top-K influential nodes in mobile social networks. In: **Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10**. New York, New York, USA: ACM Press, 2010. p. 1039. Disponível em: <<http://dl.acm.org/citation.cfm?id=1835804.1835935>>. Citado nas páginas 112 e 113.

WASSERMAN, S.; FAUST, K. **Social Network Analysis: Methods and Applications**. [S.l.]: Cambridge University Press, 1995. Structural Analysis in the Social Sciences(No. 8). 857 p. Citado na página 37.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. **Nature**, v. 393, n. 6684, p. 440–442, 1998. Citado nas páginas 13, 19, 20, 27, 42, 46, 77 e 98.

_____. Collective dynamics of 'small-world'. **Nature**, v. 393, p. 440–442, 1998. Citado na página 91.

WENG, L.; MENCZER, F.; AHN, Y.-Y. Virality prediction and community structure in social networks. **Scientific reports**, Nature Publishing Group, v. 3, p. 2522, jan 2013. ISSN 2045-2322. Disponível em: <<http://www.nature.com/srep/2013/130828/srep02522/full/srep02522.html>>. Citado nas páginas 19, 110, 112 e 121.

WINGFIELD, M. I. N. Pokemon go brings augmented reality to a mass audience. July, 2016. Disponível em: <<http://www.nytimes.com>>. Citado na página 28.

WORRELL, J.; WASKO, M.; JOHNSTON, A. Social network analysis in accounting information systems research. **Int J Account Inf Syst** (2011), n. 0, p. –, 2011. ISSN 1467-0895. InPress. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1467089511000388>>. Citado na página 37.

WU, F.; HUBERMAN, B. Novelty and collective attention. In: **Proc Nat Acad Sci USA**. [S.l.: s.n.], 2007. p. 104:17599–601. Citado na página 79.

Xulvi-Brunet, R.; Sokolov, I. M. Changing Correlations in Networks: Assortativity and Dissortativity. **Acta Physica Polonica B**, v. 36, p. 1431, may 2005. Citado nas páginas 15, 78, 122 e 123.

ZANETTE, D. H. Critical behavior of propagation on small-world networks. **Phys. Rev. E**, v. 64, n. 050901, p. 050901, 2001. Citado nas páginas 27, 28 e 77.

_____. Dynamics of rumor propagation on small world networks. **Phys. Rev. E**, v. 64, p. 050901, 2001. Citado nas páginas 77 e 121.

ZHANG, X.; ZHU, J.; WANG, Q.; ZHAO, H. Identifying influential nodes in complex networks with community structure. **Knowledge-Based Systems**, v. 42, p. 74–84, apr 2013. ISSN 09507051. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0950705113000294>>. Citado na página 121.

ZHAO, L.; WANG, Q.; CHENG, J.; CHEN, Y.; WANG, J.; HUANG, W. Rumor spreading model with consideration of forgetting mechanism: A case of online blogging livejournal. **Physica A: Statistical Mechanics and its Applications**, v. 390, n. 13, p. 2619 – 2625, 2011. ISSN 0378-4371. Citado nas páginas 79 e 80.