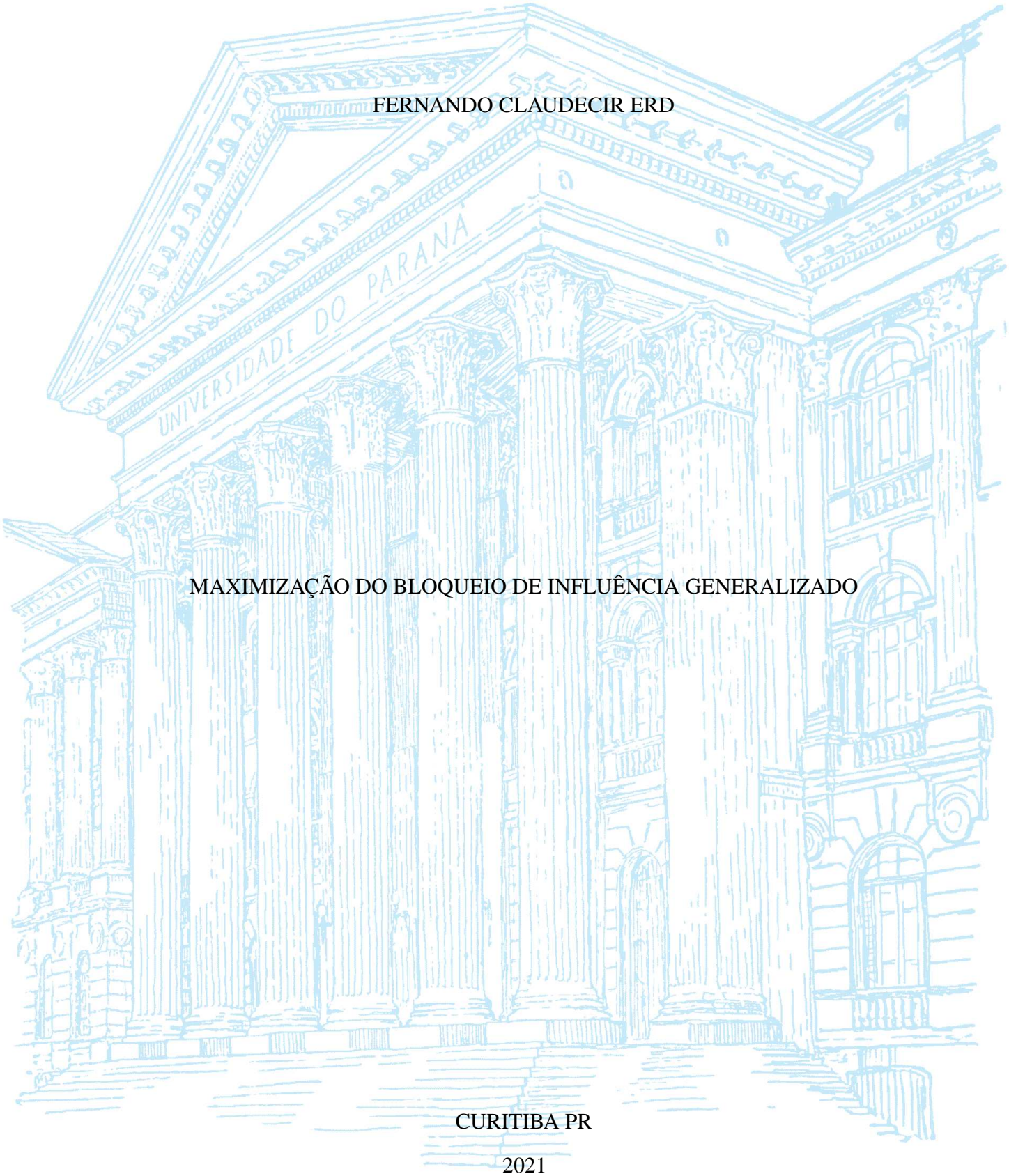
UNIVERSIDADE FEDERAL DO PARANÁ

FERNANDO CLAUDECIR ERD



MAXIMIZAÇÃO DO BLOQUEIO DE INFLUÊNCIA GENERALIZADO

CURITIBA PR

2021

FERNANDO CLAUDECIR ERD

MAXIMIZAÇÃO DO BLOQUEIO DE INFLUÊNCIA GENERALIZADO

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Informática no Programa de Pós-Graduação em Informática, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientadores: Prof. Dr. André L. Vignatti e Prof. Dr. Murilo V. G. da Silva.

CURITIBA PR

2021

CATALOGAÇÃO NA FONTE – SIBI/UFPR

E66m Erd, Fernando Claudecir

Maximização do bloqueio de influência generalizado [recurso eletrônico]/ Fernando Claudecir Erd - Curitiba, 2021.

Dissertação apresentada ao curso de Pós-Graduação em Informática, Setor de Ciências Exatas, da Universidade Federal do Paraná. Área de concentração: Ciência da Computação.
Orientadores: Prof. Dr. André L. Vignatti
Prof. Dr. Murilo V. G. da Silva.

1. Gerenciamento da informação. 2. Tecnologia da informação. I. Vignatti, André L. II. Silva, Murilo V. G. da. III. Título. IV. Universidade Federal do Paraná.

CDD 658.4038

Bibliotecária: Vilma Machado CRB9/1563

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **FERNANDO CLAUDECIR ERD** intitulada: **Maximização do Bloqueio de Influência Generalizado**, sob orientação do Prof. Dr. ANDRÉ LUÍS VIGNATTI, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 12 de Março de 2021.

Assinatura Eletrônica
16/03/2021 10:54:36.0
ANDRÉ LUÍS VIGNATTI
Presidente da Banca Examinadora (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
17/03/2021 16:07:24.0
ELIAS PROCÓPIO DUARTE JÚNIOR
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
16/03/2021 18:24:22.0
JAIME COHEN
Avaliador Externo (UNIVERSIDADE ESTADUAL DE PONTA GROSSA)

Dedico aos meus pais.

AGRADECIMENTOS

Agradeço inicialmente aos meus orientadores André Vignatti e Murilo da Silva, que aceitaram me ajudar nesse desafio, sempre disponibilizando atenção quando precisei, seja nas reuniões virtuais ou presenciais, até conversas nos corredores do departamento de informática.

Agradeço aos meus pais Cleusa e Claudecir que sempre com muito carinho e apoio, não mediram esforços para que eu chegasse até esta etapa da minha vida. Além disso, agradeço aos meus avós Teodoro, Noeli, Aloise e Carmelinda. Meus padrinhos Joel e Eva, assim como meus tios e tias e primos, em especial meu primo Joel.

Agradeço especialmente a Alexandre, Cândido, Giovanne, Hamer, Jedian, Ligia, Maíra, Ricardo e Strozzi que sempre estiveram por perto para ouvir meus desabafos e sempre me dando motivação para continuar com o trabalho.

Agradeço ao C3SL e toda equipe do Dados Educacionais, em especial Picussa, João, Henrique, Pedro e Didonet com quem sempre estávamos toda manhã no laboratório desenvolvendo a plataforma. Além de outras pessoas e professores do C3SL.

Agradeço aos demais amigos que fiz durante a graduação em ciência da computação, como André, Ruschel, Rudolf, Vytor, Gustavo, Lucas, Marcela, Aline, Tissei, Meyer, Davisson, Barreto, Sayuri, Sthephanie, entre outras pessoas com quem convivi nesses espaços durante a graduação.

Agradeço ao grupo Teoria, em especial Renato que sempre foi atencioso em tirar dúvidas sobre alguns assuntos tratados neste trabalho.

Agradeço aos meus professores da graduação que me forneceram bases sólidas para a conclusão deste trabalho, em especial Renato, Castilho, Roberto, Maziero, Elias, Menotti entre outros que não mediram esforços para ensinar da melhor maneira possível.

Agradeço a Universidade Federal do Paraná pela oportunidade de passar 6 anos da minha vida neste lugar que consegui me desenvolver como pessoa.

RESUMO

O termo desinformação pode ser entendido como uma informação falsa, dada no propósito de confundir ou induzir a erro. A partir dessa definição é possível relacionar o termo com o problema de *maximização do bloqueio de influência*. Esse problema é definido como: dada uma rede e um conjunto de vértices que são os pontos de partida para a disseminação de uma desinformação e um inteiro k, seu objetivo é encontrar k vértices na rede para serem pontos de partida para uma informação concorrente, de modo que o alcance da desinformação seja minimizado. Esse problema está altamente relacionado a disseminação de notícias falsas, abordá-lo é uma das maneiras de frear a disseminação de notícias falsas em redes sociais. Nos trabalhos da literatura, não são levados em consideração um custo para a escolha dos vértices. Tendo isso em vista, propomos cenários diferentes para esse problema, atribuindo custos diferentes para os vértices da rede, onde há um "orçamento" para escolher vértices da solução, chamamos esse problema de *maximização do bloqueio de influência generalizado*. Apresentamos demonstrações de propriedades matemáticas para o problema que podem garantir uma aproximação com um algoritmo guloso em relação à solução ótima. Ademais, são realizados experimentos que mostram que o sucesso de uma determinada estratégia varia substancialmente, dependendo da função que determina o custo de cada vértice. Em particular, investigamos a função de custo implicitamente usada em trabalhos anteriores na área que chamamos de *custo uniforme* (ou seja, todos os vértices têm custo 1) e uma função de custo que atribui custos de acordo com o grau dos vértices, chamada de *penalização de grau*. Mostramos que, embora as estratégias com bom desempenho nesses dois casos sejam muito diferentes umas das outras, ambas se correlacionam bem com estratégias simples de medidas de centralidade.

Palavras-chave: Maximização do Bloqueio de Influência, Disseminação de Informações, Redes Sociais.

ABSTRACT

The term disinformation can be understood as false information, either to confuse or mislead. From this definition, it is possible to relate the term to the problem of *influence blocking maximization*. This problem is defined as follows: given a network and a set of vertices that are the starting points for the dissemination of misinformation and an integer k, your goal is to find the k vertices in the network to be starting points for concurrent information so that the scope of disinformation is minimized. This problem is highly related to the spread of fake news, addressing it is one of the ways to stop the spread of fake news on social networks. In the literature, the cost for choosing the vertices is not taken into account. With this in mind, we propose different scenarios for this problem, assigning different costs to the vertices of the network, where there is a "budget" to choose vertices of the solution, we call this problem *generalized influence blocking maximization*. We present demonstrations of mathematical properties for the problem that can guarantee an approximation with a greedy algorithm concerning the optimal solution. Besides, experiments are carried out that show that the success of a given strategy varies, depending on the function that determines the cost of each vertex. In particular, we investigate the cost function implicitly used in previous work in the area we call *uniform cost* (that is, all vertices have a cost of 1) and a cost function that assigns costs according to the degree of the vertices, called *degree penalty*. We show that, although the strategies with good performance in these two cases are very different from each other, both correlate well with simple strategies of measures of centrality.

Keywords: Influence Blocking Maximization, Information Spread, Misinformation.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A disseminação de informações falsas não é um fenômeno novo, no entanto, com o uso crescente de redes sociais online, esse problema está ganhando mais força, como aborda Lazer et al. (2018) [18]. Segundo Lewandowsky et al. (2012) [21] existem evidências de que as pessoas tendem a acreditar em informações que correspondem à sua percepção das narrativas sociais e a desacreditar nas narrativas que desconstroem essa percepção. Dessa forma, as mídias sociais, devido à sua estrutura e formas de disseminação de informações, poderiam expandir a circulação de desinformações. Abordando plataformas mais usuais de redes sociais Vosoughi et al. (2018) [30] mostram que na rede social Twitter as chances de uma notícia falsa ser compartilhada são 70% maiores do que as de compartilhar uma notícia real.

Recentemente, vários estudos mostraram como a disseminação de informações falsas tem potencial para impactar no comportamento da sociedade. Um dos trabalhos relevantes é o de Allcott e Gentzkow (2017) [1], que apresenta uma análise de como a desinformação pode ter afetado o resultado das eleições presidenciais de 2016 nos Estados Unidos. Outro exemplo é o número de fontes questionáveis nas principais plataformas sociais em relação ao surto do vírus COVID-19, como mostra Cinelli et al. (2020) [7].

Em contrapartida, surge a questão sobre se é possível evitar que a desinformação tenha um grande impacto em diversas áreas. Atualmente diversas agências de comunicação estão criando equipes de checagem de informação como alternativa ao combate, pode-se citar a Agência Lupa, da Folha de S. Paulo (2021) [9] e "Fato ou Fake" do Grupo Globo (2021) [14]. Além disso, pesquisadores da USP (2018) [10] criaram um sistema capaz de detectar notícias falsas utilizando aprendizado de máquina. Uma alternativa ao combate da desinformação é a divulgação massiva de notícias de veículos de informação confiáveis, como pode ser conferido na reportagem da Folha de S. Paulo (2021) [28] com o título "Disseminação do jornalismo profissional reduz influência de fake news, indica pesquisa". O trabalho que apresentamos pode ser entendido como uma abordagem matemática e computacional onde o combate de disseminações negativas pode ser evitado (ou combatido) com a divulgação de disseminações positivas.

Os aspectos algorítmicos de um problema originalmente do campo do "marketing viral" foram investigados por Kempe et al. (2003) [17]. O problema computacional proposto, conhecido como *maximização de influência* em redes, é informalmente definido como: dada uma rede em que um vértice corresponde a uma pessoa e uma aresta corresponde à conexão entre duas pessoas, o objetivo é selecionar os melhores indivíduos para anunciar um produto, para que as informações sobre este produto atinjam o maior número de pessoas possíveis na rede, assim maximizando o alcance. A partir desse problema, surgiu uma linha de pesquisa que procura encontrar uma contra-estratégia para limitar a disseminação de uma influência, proposto por He et al. (2011) [16]. Neste trabalho iremos assumir que estamos lidando com a disseminação de informações falsas e necessitamos de uma estratégia que busca espalhar as informações corretas (ou uma contra-narrativa) de maneira efetiva pela rede. Esse problema computacional, chamado de *maximização do bloqueio de influência*, é definido informalmente da seguinte maneira. Dado um conjunto de vértices como ponto de partida para a disseminação de informações falsas na rede e um número inteiro k, o objetivo é encontrar um conjunto de vértices de tamanho k para iniciar a disseminação de informações corretas pela rede, tendo como objetivo minimizar o alcance da informação incorreta.

1.1 OBJETIVOS

O objetivo deste trabalho é contribuir com a pesquisa na área de disseminação de informações e serão listados em tópicos a seguir.

- Levar ao leitor fundamentos teóricos sobre modelos de propagação de informação, mais especificamente, os modelos de cascata independente de múltiplas campanhas, cascata independente de campanha inconsciente ambos utilizados por Budak et al. (2011) [5] e limitante linear competitivo proposto por He et al. (2011) [16].
- Apresentar a definição do problema *maximização do bloqueio de influência* e um semelhante chamado *limitação eventual de influência*, assim como propriedades de ambos os problemas.
- Expor o contexto histórico de como a pesquisa evoluiu na área, até chegar no presente.
- Propor o problema *maximização do bloqueio de influência generalizado*.
- Abordar propriedades matemáticas que garantem um fator de aproximação para o problema proposto.
- Discutir resultados usando como estratégia medidas de centralidade de grafos com o intuito de diminuir o alcance de uma informação negativa.

1.2 METODOLOGIA

Os experimentos e resultados que serão apresentados neste trabalho mostram como o problema *maximização do bloqueio de influência generalizado* se comporta em doze cenários diferentes para grafos do mundo real. Para cada grafo foram realizados experimentos em grafos direcionados, em seguida as direções foram ignoradas para analisar as diferenças no comportamento. Além disso, verificamos o problema em diferentes cenários de propagações, no qual temos a probabilidade da disseminação avançar para outros vértices variando em um certo intervalo preestabelecido. Os experimentos mostram que a disseminação sobre o modelo cascata independente de multi campanhas (mais detalhes na Seção 2.1.1) com medidas de centralidades complexas, como percolação e intermediação, apresentam um comportamento similar à estratégia simples de escolher os vértices de graus maiores em grafos do mundo real. Além disso, propomos uma generalização mais realista do problema *maximização do bloqueio de influência*, onde é acrescentada uma função de custo como entrada do problema. Essa função representa o custo de escolher cada vértice v na rede, tornando-o mais real do nosso ponto de vista, o qual chamamos de *maximização do bloqueio de influência generalizado*. Parte dos resultados apresentados nesta dissertação foram publicados em *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* [11].

1.3 ORGANIZAÇÃO

O trabalho está dividido da seguinte maneira: o Capítulo 2 detalha os fundamentos teóricos sobre modelos de disseminação, retrata os problemas *limitação eventual de influência* e *maximização do bloqueio de influência* com detalhes, apresentando algoritmos gulosos propostos na literatura e apresenta os conceitos de funções submodular e monotônica. A bibliografia quando tratado o problema de *maximização do bloqueio de influência* e similares é apresentada

no Capítulo 3. O problema que propomos é apresentado no Capítulo 4, assim como propriedades que exploramos. A metodologia para avaliação experimental da proposta e os resultados do trabalho são apresentados no Capítulo 5. Por último, no Capítulo 6 tratamos das considerações finais do trabalho.

2 FUNDAMENTOS

Nesse capítulo, são apresentados os modelos de cascata independente de múltiplas campanhas, cascata independente de campanha inconsciente e limitante linear competitivo que são usados pela literatura para simular a propagação de informações em um grafo. Em seguida são detalhados os problemas computacionais de *limitação eventual de influência* e *maximização do bloqueio de influência*. E por fim é apresentado o conceito de submodularidade.

2.1 MODELOS DE DISSEMINAÇÃO DE INFORMAÇÕES

O processo que descreve como a disseminação de informações ocorre pela rede de usuários em mídias e redes sociais é estudado de forma vasta. Kempe et al. (2003) [17] propôs dois modelos amplamente usados pela comunidade para simular o comportamento da propagação de informações, sendo estes chamados de cascata independente e limitante linear. Esse processo pode ser utilizado em um grafo que representa, por exemplo, uma rede de encaminhamento de notícias, onde uma aresta direcionada de v para w indica que o usuário representado pelo vértice v pode encaminhar uma informação, verdadeira ou falsa, para o usuário do vértice w.

O modelo de cascata independente é baseado em probabilidades, e de maneira geral, associa a cada aresta uma certa probabilidade de transmitir uma informação. Por outro lado, no modelo limitante linear a disseminação é espalhada pela rede tendo como base a quantidade de vizinhos de um vértice v que adotam a informação, ou seja, quanto mais vizinhos de um vértice v aceitarem uma informação, mais propenso o vértice v está em adotá-la. Quando tratamos de mais de uma informação propagando-se pela rede, Budak et al. (2011) [5] propuseram duas variações do modelo de cascata independente, chamados de cascata independente de multi campanha e cascata independente de campanha inconsciente. De maneira semelhante, He et al. (2011) [16] apresentaram uma versão do modelo limitante linear para duas disseminações propagando-se pela rede, chamado de limitante linear competitivo. Como ambos modelos apresentam várias definições em comum, inicialmente denotaremos suas semelhanças e posteriormente aprofundaremos os detalhes específicos para cada um.

O primeiro ponto em comum nos modelos apresentados neste trabalho é que atuam em um grafo direcionado. Além disso, sejam N e P duas cascatas de informações na rede, sendo que N modela o processo de disseminação da desinformação pela rede e P representa a informação verdadeira. As cascatas tem seu início a partir de um conjunto de vértices, esses conjuntos que iniciam a propagação denotados por N_0 e S, sendo N_0 o conjunto que inicia a propagação da cascata negativa e S o conjunto de vértices que dão início a disseminação da informação verdadeira. É comum na literatura da área nomear os conjuntos N_0 e S como o conjunto de “sementes"do início da propagação.

Cada vértice na rede pode apresentar 3 estados diferentes, sendo eles: positivo, negativo e inativo. O estado positivo para um vértice v descreve que ele aceitou a informação verdadeira. De maneira semelhante o estado negativo representa que o vértice aceitou a desinformação. Caso um vértice não aceite nenhuma informação, tem seu estado denominado como inativo. Inicialmente, os vértices do conjunto N_0 são atribuídos como negativos, os vértices do conjunto S como positivos, e o restante dos vértices do grafo são atribuídos como inativos.

Uma etapa t dos modelos de informações consiste em todos os vértices que tornaram-se ativos com a informação positiva ou negativa na etapa $t-1$ tentarem passar a informação para seus vizinhos. No caso de grafos direcionados seus vizinhos de saída. Além disso, cada vértice

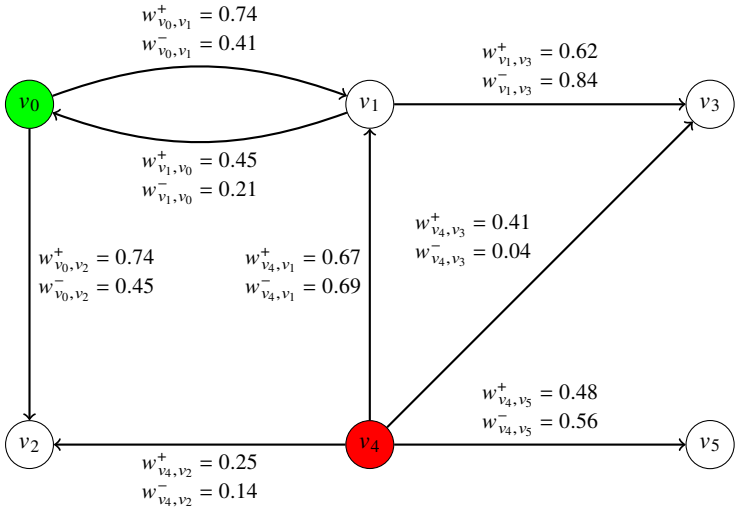


Figura 2.1: Grafo no estado inicial, ou seja, $t = 0$, onde $S = \{v_0\}$, $N_0 = \{v_4\}$, que representa o conjunto de sementes positivas e negativas respectivamente.

tem uma única chance de enviar a informação para seus vizinhos durante o processo. Se as cascatas P e N tentarem ativar um vértice ao mesmo tempo, vale a pena ressaltar que no trabalho original de Budak et al. (2011) [5], a informação positiva tem preferência. Entretanto, como a maioria dos trabalhos posteriores deram prioridade para a informação negativa, levaremos em consideração que a informação negativa tem precedência. Sendo assim, a cascata N tem prioridade em caso de "empate". Por último, depois que um vértice altera seu estado, para positivo ou negativo, o seu estado permanece inalterado até o fim da simulação das disseminações. O processo é executado até que não haja mais nenhuma ativação disponível para uma nova etapa.

2.1.1 Cascata Independente de Multi Campanhas (MCICM)

O modelo de cascata independente de multi campanhas (em inglês *Multi-Campaign Independent Cascade Model (MCICM)*), prevê duas informações sendo disseminadas simultaneamente em uma rede de maneira que cada uma se propaga, pelas arestas, de forma diferente em relação a outra. Nesse modelo, cada aresta possui dois pesos, w^+ e w^- que atuam como probabilidades de envio de informação de um vértice a outro, assim como sua aceitação, ou seja, o peso w^+_{uv} da aresta é a probabilidade do vértice u ativar o vértice v, assim transmitindo a informação positiva para o vértice v. De maneira análoga w^-_{uv} é a probabilidade do vértice u ativar o vértice v, assim transmitindo a informação negativa para o vértice v.

Um exemplo de como a informação é transmitida em um grafo pode ser observado nas Figuras 2.1, 2.2 e 2.3, considerando $N_0 = \{v_4\}$ como o conjunto das sementes negativas e $S = \{v_0\}$ como o conjunto das sementes positivas. O primeiro passo é alterar o estado dos vértices pertencentes aos conjuntos S e N_0 para positivo e negativo respectivamente. Como a informação negativa tem prioridade, os vértices pertencentes a N_0 tentam ativar seus vizinhos, sendo assim o vértice v_4 faz uma tentativa de envio de informação falsa para seus vizinhos de saída, ou seja, os vértices v_1, v_2, v_3 e v_5. Como resultado na Figura 2.2 o vértice v_1 acaba sendo influenciado pela informação falsa iniciada de v_4, e como a tentativa é única, v_4 não terá outra chance de enviar a informação para os vértices que ele falhou em tentar influenciar com a informação. Com a transmissão negativa finalizada na etapa $t = 1$, a cascata de informações positivas (que inicia pelo vértice v_0) começa a disseminação, tentando contaminar seus vizinhos inativos, neste caso apenas v_2, uma vez que v_1 já foi contaminado com a informação falsa. A

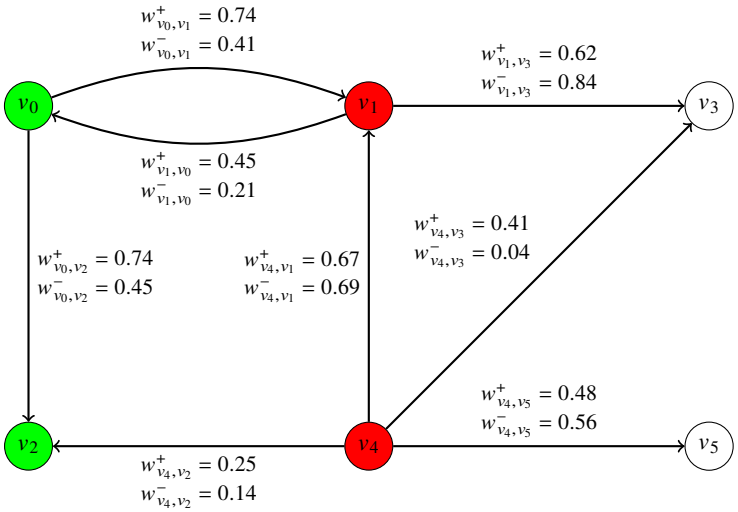


Figura 2.2: Estado do grafo na etapa $t = 1$, nessa etapa o conjunto de sementes negativas N_0 consegue convencer o vértice v_1 que a notícia falsa remete a uma verdade, entretanto falha em convencer v_2, v_3 e v_5. Enquanto que o vértice de origem da contra informação v_0 consegue convencer o vértice v_2. Lembrando que $N_0 = \{v_4\}$ e $S = \{v_0\}$.

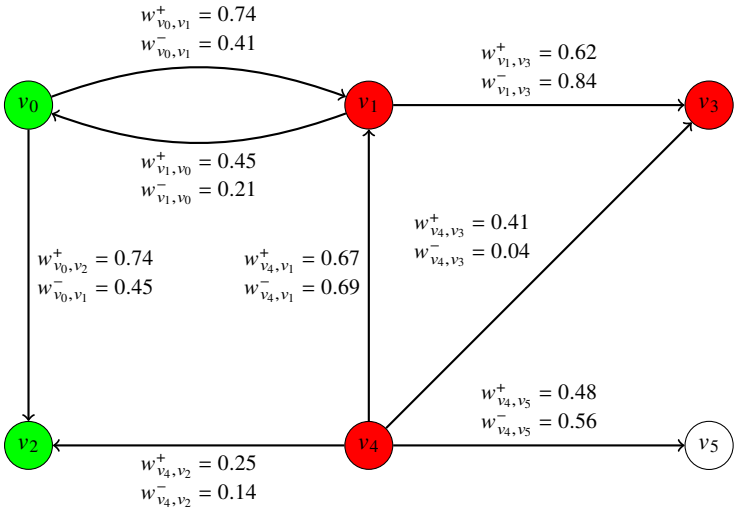


Figura 2.3: Grafo G na etapa $t = 2$, temos que os vértices ativados na última etapa (v_2 e v_3) tem a chance de passar a informação para seus vizinhos, neste caso v_3 ativa com informação negativa o vértice v_4 e a cascata positiva para, pois o vértice v_2 não tem vizinhos para repassar a informação.

Figura 2.2 mostra o estado do grafo após a etapa $t = 1$, com v_2 sendo contaminado positivamente pelo vértice v_0.

De maneira similar, o processo continua, já que na próxima etapa vértices que se tornaram negativos na etapa anterior tentam contaminar seus vizinhos inativos. De maneira análoga o processo ocorre para os vértices positivos, ou seja, o vértice v_1 tem a possibilidade de contaminar o vértice v_3. Entretanto, a cascata positiva é paralisada, já que o vértice v_2 não possui vizinhos de saída. A Figura 2.3 mostra o grafo com o vértice v_1 ativando com sucesso o vértice v_3, neste momento acaba a simulação já que ambas as cascatas não conseguem ativar nenhum outro vértice.

2.1.2 Cascata Independente de Campanha Inconsciente (COICM)

O modelo de cascata independente de campanha inconsciente (em inglês *Campaign-Oblivious Independent Cascade*) é usado quando tratamos de duas disseminações com características semelhantes de propagação. Neste caso é definida uma única probabilidade em cada aresta, ou seja, tanto a informação positiva como a negativa têm as probabilidades de propagação iguais em todas as arestas. Esta versão alternativa representa uma disseminação onde as informações têm propagação idêntica em cada aresta. O restante do modelo é similar ao de MCICM.

2.1.3 Limitante Linear Competitivo (CLT)

O limitante linear competitivo é uma extensão do modelo determinístico de disseminação LT (*linear threshold*), proposto por Kempe et al. (2003) [17]. Diferentemente dos modelos cascatas que são probabilísticos, este modelo trata o processo de disseminação sob a visão determinística. A versão com duas disseminações do modelo proposto por He et al. (2011) [16] é a seguinte.

Cada vértice v tem dois limitantes entre 0 e 1: um representa o limitante positivo, denotado por θ_v^+, e outro o limitante negativo, denotado por θ_v^-. De maneira similar aos modelos de cascata, cada aresta possui dois pesos $w_{u,v}^+$ e $w_{u,v}^-$, que nesse caso indicam o peso de influência positiva e negativa do vértice u em relação a v. Além disso, seja $N(u)$ o conjunto de vértices vizinhos de entrada do vértice u, então o modelo deve respeitar as seguintes condições:

$$\sum_{v \in N(u)} w_{v,u}^+ \leq 1$$
$$\sum_{v \in N(u)} w_{v,u}^- \leq 1$$

ou seja, de forma sucinta, podemos dizer que a soma dos pesos dos vizinhos de entrada de u não pode ultrapassar 1, e a soma pertence ao intervalo [0,1].

Na etapa 0, existem dois conjuntos de sementes separados, o conjunto de sementes positivas S e o conjunto de sementes negativas N_0. Em cada etapa t, a informação positiva e a informação negativa se propagam independentemente, assim como no modelo limitante linear original, usando pesos/limitantes positivos e pesos/limitantes negativos, respectivamente. Dado um vértice v, ele torna-se positivo se a soma dos pesos das arestas dos seus vizinhos de entrada ativados positivamente for maior que seu limitante θ_v^+, ou seja

$$\sum_{w \in N^S(v)} w^+_{w,v} \geq \theta^+_v$$

onde $N^S(v)$ é o conjunto de vizinhos de entrada de v ativados positivamente. De maneira análoga, temos que v é ativado pela disseminação negativa se a soma dos pesos das arestas do seus vizinhos de entrada ativados negativamente for maior que seu limitante θ^-_v, ou seja

$$\sum_{w \in N^a(v)} w^-_{w,v} \geq \theta^-_v$$

onde $N^a(v)$ é o conjunto de vizinhos de entrada de v ativados negativamente.

2.2 PROBLEMAS

Nesta seção, abordaremos a definição de dois problemas diferentes. O primeiro é o de *limitação eventual de influência*, que trata sobre a disseminação de informações em cascatas, proposto inicialmente por Budak et al. (2011) [5]. A seguir detalhamos o problema sobre a perspectiva do modelo limitante linear competitivo, sendo chamado de *maximização do bloqueio de influência*.

2.2.1 Limitação Eventual de Influência (EIL)

O problema *limitação eventual de influência* (em inglês *eventual influence limitation problem*) tem como foco conter a disseminação de informações falsas pela rede, ou seja, o objetivo é minimizar o número de pessoas que passam a acreditar em uma desinformação.

Sejam N e P as cascatas de informações que representam as informações negativa e positiva, respectivamente, assim como os conjuntos S e N_0, que denotam os vértices que iniciam a disseminação de informações positiva e negativa. Usaremos I_N para denotar o conjunto de vértices que a cascata N consegue influenciar na ausência da cascata P, iniciando pelos vértices do conjunto N_0. Além disso, seja $\pi\colon 2^V \to \mathbb{N}$ uma função tal que $\pi(S)$ representa o tamanho do subconjunto de I_N em que a cascata P iniciando pelos vértices de S consegue evitar que adotem a informação negativa, ou seja, a quantidade de vértices que se tornam positivos ou permanecem inativos do conjunto I_N. Então o problema de *limitação eventual de influência* é equivalente à seleção de um conjunto S, de modo que a esperança de $\pi(S)$ seja maximizada.

Para a entrada do problema, os autores consideram que existe um único adversário n_a, ou seja, $N_0 = \{n_a\}$. Além disso, os autores acrescentam um atraso r, que significa a quantidade de etapas de uma simulação em que a cascata positiva P leva para iniciar o processo de disseminação de informação positiva, e um inteiro k, que representa o tamanho do conjunto S.

Como o modelo de cascata independente é um processo estocástico, para calcular experimentalmente π para um determinado conjunto de vértices, é necessário executar várias simulações para obter uma estimativa do conjunto esperado. Com isso, propuseram o Algoritmo 1, que usa uma abordagem gulosa para obter uma aproximação (mais detalhes na Seção 2.3). O algoritmo tem seu início definindo o conjunto S igual a vazio e define o número de iterações como 10000 (variável R). Temos que para cada vértice é calculado o procedimento $InfLimit(n_a, r, S, v)$ que consiste em primeiro atribuir pesos aleatórios para cada aresta no grafo e simular a limitação

de influência, dado que o adversário é o vértice n_a, a cascata negativa é detectada com atraso r, o conjunto de vértices que já optaram por ativar inicialmente na cascata P é S e o vértice do qual estamos avaliando a influência é v. Esse método retorna o ganho marginal da adição do vértice v ao conjunto S, ou seja, o número de vértices que $S \cup \{v\}$ pode salvar e o conjunto S não pode. Por fim, temos a divisão s_v pela quantidade de iterações R, desse modo obtendo o valor esperado do ganho marginal quando analisado o vértice v.

Algoritmo 1: Algoritmo Guloso para Limitação Eventual de Influência

Entrada: Dado n_a, r, k onde n_a denota o adversário inicial, r o atraso em que a cascata N é detectada e k o número de vértices inicialmente ativos em S
Saída: Um conjunto de sementes S

```
1  início
2      S = ∅
3      R = 10000
4      para i = 1 até k faça
5          para cada vértice v ∈ V − S faça
6              s_v = 0
7              para j = 1 até R faça
8                  s_v += InfLimit(n_a, r, S, v)
9              fim
10             s_v = s_v/R
11         fim
12         // Escolhe o vértice i que maximiza π(S ∪ {i}) − π(S)
           S = S ∪ {argmax_{v∈V\S} {s_v}}
13     fim
14 fim
15 retorna S
```

2.2.2 Maximização do Bloqueio de Influência (IBM)

O problema de *limitação eventual de influência* forneceu bases para He et al. (2011) [16] proporem o problema de *maximização do bloqueio de influência* sobre o modelo de disseminação limitante linear competitivo. A definição do problema é a seguinte.

Sejam θ^+ e θ^- os vetores dos limitantes positivos e negativos do modelo limitante linear competitivo, respectivamente. Assim, definimos $IBS(S, N_0|\theta^+, \theta^-)$ como o conjunto dos vértices v em G que sob limitante dos vetores θ, v é ativado como negativo se N_0 for o conjunto de sementes negativas e o conjunto S de sementes positivas for igual a vazio. Além disso, v não é ativado negativamente se o conjunto S de sementes positivas não for vazio. De maneira mais sucinta, podemos dizer que $IBS(S, N_0|\theta^+, \theta^-)$ é o conjunto de vértices que sem uma contra-estratégia tornam-se negativos, entretanto ficam positivos ou inativos quando a contra-estratégia para limitar a informação é executada a partir de um conjunto de sementes positivas S.

Definimos $\sigma_{NIR}(S)$ como a *redução de influência negativa* de um conjunto de sementes positivas S, como o valor esperado, sobre todos θ^+ e θ^-, do tamanho de $IBS(S, N_0|\theta^+, \theta^-)$, ou seja:

$$\sigma_{NIR}(S) = E_{\theta^+\theta^-}(|IBS(S, N_0|\theta^+, \theta^-)|) \tag{2.1}$$

então, pode-se dizer que *maximização do bloqueio de influência* é o problema de encontrar um conjunto de sementes positivas S de tamanho máximo k, que maximiza $\sigma_{NIR}(S)$, isto é

$$S* = \text{argmax}_{|S|\leq k}\sigma_{NIR}(S) \tag{2.2}$$

Algoritmo 2: ALGORITMO GULOSO

Entrada: Um grafo $G = (V, E)$, e $k \in \mathbb{N}$
Saída: Um conjunto de sementes S, de modo que $S \subseteq V \setminus N_0$

```
início
    S = ∅
    para i = 1 até k faça
        Seleciona u = argmax_{v∈V\(N0∪S)}(σ_NIR(S ∪ {v}))
        S = S ∪ {u}
    fim
fim
retorna S
```

De maneira similar, os autores propuseram um algoritmo guloso que garante uma aproximação da solução ótima. O Algoritmo 2 inicialmente define o conjunto S vazio, a rodada inicial escolhe o vértice que maximiza σ_{NIR} e o adiciona em S. Nas próximas $k - 1$ rodadas é selecionado o vértice que unido a S maximiza σ_{NIR}, resultando assim em um conjunto de tamanho k.

2.3 PROPRIEDADES DA FUNÇÃO DE INFLUÊNCIA

Para obter uma aproximação garantida de solução ótima para os problemas *limitação eventual de influência* e *maximização do bloqueio de influência* é necessário provar que a função de influência σ de cada problema apresenta a propriedade de ser submodular e monotônica, que são descritas a seguir.

Seja U um conjunto e seja f uma função arbitrária definida como $f : 2^U \rightarrow \mathbb{R}^+$. Dizemos que f é *submodular* se possui a propriedade de diminuição de ganho, isto é, se o ganho obtido ao adicionar um determinado elemento a um conjunto S é pelo menos tão alto quanto adicionar o mesmo elemento a um superconjunto de S, como mostra a Definição 1.

Definição 1: *Sejam S, T e U conjuntos tais que* $S \subseteq T \subseteq U$ *e* $f : 2^U \rightarrow \mathbb{R}^+$. *Dizemos que* f *é* submodular *se* $f(S \cup \{w\}) - f(S) \geq f(T \cup \{w\}) - f(T)$ *para todo* $w \in U \setminus T$.

A mesma função f também é *monotônica*, quando o ganho marginal da função é não decrescente, como mostra a Definição 2.

Definição 2: *Sejam S, T, e U conjuntos tais que* $S \subseteq T \subseteq U$ *e* $f : 2^U \rightarrow \mathbb{R}^+$. *Dizemos que* f *é* monotônica *(não decrescente) se* $f(S) \leq f(T)$.

Foi provado por Cornuejols et al. (1977) [8] e Nemhauser e Fisher (1978) [13] que dada uma função f, se f é *submodular* e *monotônica*, existe um algoritmo guloso baseado no princípio de *escalada de montanha* que garante um fator de aproximação $1 - \frac{1}{e}$ em relação a solução ótima. Em outras palavras, pode-se dizer que escolher o vértice de maior ganho marginal em todo passo, acaba oferecendo uma aproximação de $1 - \frac{1}{e}$ da solução ótima.

Neste capítulo aprofundamos em entender como é possível simular a disseminação de informações sobre um olhar matemático e computacional, adquirindo bases para entender dois problemas fundamentais nessa área de pesquisa. Dessa forma é possível compreender como o conceito de submodularidade é usado para conseguir avanços que serão detalhados no próximo capítulo onde é realizado uma revisão sobre os trabalhos relacionados.

3 REVISÃO BIBLIOGRÁFICA

A partir do problema da *maximização de influência* de Kempe et al. (2003) [17] surgiram várias ramificações da linha de pesquisa, uma delas é quando trata-se do comportamento de uma disseminação em contra-atacar outra. O primeiro trabalho relacionado foi o de Budak et al. (2011) [5], que propôs o problema *limitação eventual de influência* (EIL) detalhado na Seção 2.2.1. A principal contribuição dos autores é a demostração de que o problema apresenta função submodular e monotônica sobre o modelo COICM. Além disso, os autores mostraram que usando o modelo de cascata independente de várias campanhas (MCICM), quando as probabilidades de disseminação positiva e negativa são arbitrárias, o problema não apresenta a propriedade submodular, entretanto quando a probabilidade de disseminação positiva é igual a 1 (sendo chamada de propriedade de *alta eficácia*) para todas as arestas, o modelo garante a submodularidade e a propriedade de função monotônica, assegurando uma aproximação. Além disso, os autores demostraram que mesmo com a propriedade de alta eficácia o problema é NP-Difícil.

Sobre o modelo de limitante linear competitivo, o trabalho de He et al. (2011) [16] definiu de maneira formal o problema *maximização do bloqueio de influência* anteriormente detalhado na Seção 2.2.2 e, neste trabalho foi utilizada a estrutura de *grafo local direcionado acíclico*, e tendo como base o algoritmo LDAG [6], que aborda o problema de maximização de influência. Os autores propuseram o algoritmo CLDAG e provaram que para o modelo limitante linear competitivo o problema é NP-Difícil e fornecem a demostração de que o problema é submodular e monotônico, o que garante uma aproximação de $1 - 1/e$ da solução ótima usando uma estratégia de escalada gulosa.

A respeito dos trabalhos baseados nos modelos de cascata ocorre uma ramificação nas pesquisas, alguns trabalhos realizaram experimentos com o modelo MCICM e outros com o COICM. Para o modelo MCICM, Arazkhani et al. (2019) [3] abordaram uma métrica baseada nas medidas de centralidade de grau, intermediação e proximidade para escolher o melhor conjunto de vértices de sementes positivas. Ademais, forneceram em um trabalho posterior [2] que combinou as centralidades com um algoritmo para encontrar comunidades no grafo de entrada do problema. Inicialmente é realizado um pré-processamento para encontrar as k maiores comunidades no grafo, a partir do algoritmo *Fuzzy Clustering*, em seguida para cada comunidade é escolhido o vértice que possui o maior grau, intermediação ou proximidade que foram as métricas escolhidas pelos autores. Com relação ao modelo MCICM com a propriedade *alta eficácia* podemos relatar o trabalho de Wu e Pan (2017) [31] que utilizou a estrutura de *máxima arborescência de influência*, propondo duas heurísticas CMIA-H e CMIA-O. No mesmo trabalho foi mostrado o comportamento do CMIA sobre o modelo COICM. Além disso, os autores definiram as probabilidades das arestas como 0.2, 0.05, 0.01 nos experimentos, significando probabilidade alta, média e baixa respectivamente. Outra abordagem para escolher as probabilidades foi definir que para cada aresta (u, v), sua probabilidade é $1/d_v$, onde d_v é o grau de entrada do vértice v.

Uma versão alternativa do problema foi proposta por Zhu et al. (2018) [32], chamado de *maximização de bloqueio de influência com reconhecimento de local* (LIBM). Esse problema inclui que as informações de localização dos vértices podem desempenhar um papel importante na seleção de sementes, onde cada vértice possui uma coordenada geográfica que pode influenciar a maneira como as informações disseminam pela rede. Além disso, em trabalhos posteriores os autores adicionaram pequenas alterações na formulação do problema, propondo os problemas

seleção de sementes com base na localização para maximizar o bloqueio de influência [33] e *maximização de bloqueio de influência direcionada com reconhecimento de local* [34].

Em algumas variantes do problema de *maximização de influência*, acrescenta-se um custo de escolha para cada vértice na rede. Entre eles pode-se citar o problema de *maximização de influência com orçamento* abordado por Nguyen e Zheng (2013) [24], onde para cada vértice é associado a um custo arbitrário. O objetivo do problema é selecionar um conjunto S de modo que o custo desse conjunto seja menor ou igual a um orçamento b e que maximize o alcance na rede. Uma evolução deste problema para um contexto mais próximo com duas disseminações em uma rede é o problema *maximização da competição de influência com orçamento*. O objetivo deste problema é maximizar o alcance de um produto na rede contra um competidor. A contribuição de Pham et al. (2019) [26] foi adicionar um custo para cada vértice e demostrar que em um modelo similar ao limitante linear o problema não apresenta a propriedade submodular na função de influência.

Por fim, pode-se citar o trabalho de Leskovec et al. (2007) [20] que propuseram a formulação do problema de *detecção de surto* sob o contexto de distribuição de água. Dado um orçamento b o propósito é selecionar o melhor posicionamento de sensores para o monitoramento da qualidade de água que respeita o orçamento. Desta forma, uma contaminação começa em algum vértice e a partir do momento que a contaminação passa por um sensor, um alarme é disparado. Os autores abordam esse problema a partir três pontos de vista diferentes em seus experimentos. A probabilidade de detecção de uma contaminação, que é a fração de eventos de contaminação detectados pelos vértices selecionados para o monitoramento. O tempo de detecção da contaminação pelos sensores, ou seja, o tempo que passou do início do surto até a detecção por um dos vértices selecionados. E por último a população de vértices afetados pela contaminação, isto é, a quantidade de vértices que são salvos da contaminação pelo conjunto de sensores.

Ao concluir esse capítulo dispomos de bases sólidas para buscar compreender o contexto em que se encontra o processo de pesquisa relacionado ao problema que buscamos propor, abordando problemas similares. Um importante questionamento aqui é observar que, com algumas exceções, a grande parte dos trabalhos são recentes, então é possível concluir que a pesquisa no campo ao combate a disseminação de informações está em amplo crescimento. Agora, é viável apresentar ao leitor o problema *maximização do bloqueio de influência generalizado*, assim como explorar propriedades matemáticas sobre o problema, que serão abordadas no próximo capítulo.

4 MAXIMIZAÇÃO DO BLOQUEIO DE INFLUÊNCIA GENERALIZADO

Relatados os problemas *limitação eventual de influência* e *maximização do bloqueio de influência*, precursores da área, um dos objetivos deste trabalho é propor o problema *maximização do bloqueio de influência generalizado*, onde é acrescentada uma função de custo $c : V \to \mathbb{N}^+$ como entrada do problema, que modela o custo de escolha para os vértices.

Nesse caso, dado um grafo $G = (V, E)$ que representa uma rede complexa onde V são os vértices e E as arestas, um conjunto de sementes negativas N_0 e um k inteiro positivo, queremos encontrar o melhor conjunto de vértices pertencentes a V que tenha custo menor ou igual a k para minimizar o número de vértices que são ativados negativamente.

Seja I_N o conjunto de vértices negativos que é o resultado de uma execução do processo estocástico de disseminação, o resultado I_N depende do grafo G, das probabilidades w^+ e w^- (assumindo que trabalhamos com modelos de difusão baseados em cascata), e o conjunto de sementes iniciais negativa e positiva denotados por N_0 e S.

Assim, podemos definir que $\mathbb{E}\big[|I_N| \;\big|\; (N_0, \{\emptyset\})\big]$ é o tamanho esperado do conjunto I_N, de maneira que não existe nenhuma disseminação concorrente e isso representa o número médio de vértices que a cascata de informações negativas consegue atingir dado que não existe uma cascata positiva. Por outro lado, $\mathbb{E}\big[|I_N| \;\big|\; (N_0, S)\big]$ é o tamanho esperado do conjunto I_N, quando há disseminação positiva iniciada pelos vértices do conjunto S. Logo, queremos encontrar o conjunto S que minimize $\mathbb{E}\big[|I_N| \;\big|\; (N_0, S)\big]$.

Podemos medir o impacto de um conjunto de sementes positivas S considerando a diferença entre dois cenários, quando o conjunto de sementes positivo inicial é S e quando o conjunto de sementes positivo inicial está vazio. Isso é chamado de *influência negativa bloqueada esperada* de S e formalmente definido como

$$\sigma(S) = \mathbb{E}\big[|I_N| \;\big|\; (N_0, \{\emptyset\})\big] - \mathbb{E}\big[|I_N| \;\big|\; (N_0, S)\big].$$

Agora é possível definir o problema de *maximização do bloqueio de influência generalizado*.

Maximização do Bloqueio de Influência Generalizado (GIBM): Dado um grafo $G = (V, E)$, uma função $c : V \to \mathbb{N}^+$, as probabilidades de propagação w^+ e w^-, um conjunto de sementes negativas N_0, e um inteiro positivo k, GIBM visa encontrar o conjunto de sementes positivo S que maximiza $\sigma(S)$ onde $\sum_{v \in S} c(v) \leq k$.

O caso particular em que todos os vértices têm o mesmo peso, sem perda de generalidade $c(v) = 1$ para cada $v \in V$, é exatamente o problema de *maximização do bloqueio de influência* apresentado na Seção 2.2.2. Com o problema definido, agora é possível explorar propriedades matemáticas do GIBM. A primeira propriedade que iremos explorar é relacionada à questão de submodularidade sobre o modelo MCICM, conforme apresenta o Teorema 1.

De acordo com Kempe et al. (2003) [17], O processo de disseminação de informações para os modelos de cascata pode ser abordado da seguinte maneira: um vértice recém ativado u tentando ativar seu vizinho v com uma informação pode ser vista como um lançamento de uma moeda com o viés $w_{u,v}$. Budak et al. (2011) [5] expandiram essa observação para modelos com duas cascatas como é o caso do MCICM e COICM, ou seja, é possível ver um vértice u tentando ativar seu vizinho v com a informação negativa, como sendo uma moeda jogada com viés de $w^-_{u,v}$. De maneira análoga é possível o evento de um vértice u tentando ativar seu vizinho v com a informação positiva, como sendo uma moeda jogada com viés de $w^+_{u,v}$.

Vale a pena ressaltar que tanto faz o momento em que a moeda é lançada, se ela é lançada no momento que o vértice u tenta ativar seu vizinho v ou se foi pré-lançada e armazenado o resultado para ser avaliado no momento em que u tenta ativar v. Portanto, considerando uma instância específica de disseminação de informação, ou seja, dado um grafo G com as probabilidades $w^{+}_{u,v}$ e $w^{-}_{u,v}$ para todas as arestas é possível pré-lançar todas as moedas para determinar quais arestas do grafo G estão ativas para enviar a informação positiva e quais estão ativas para enviar a informação negativa. Desse modo é possível traçar o caminho das disseminações pela rede. A prova que apresentamos no Teorema 1 é similar ao que Budak et al. (2011) [5] demostrou para o problema *limitação eventual de influência*.

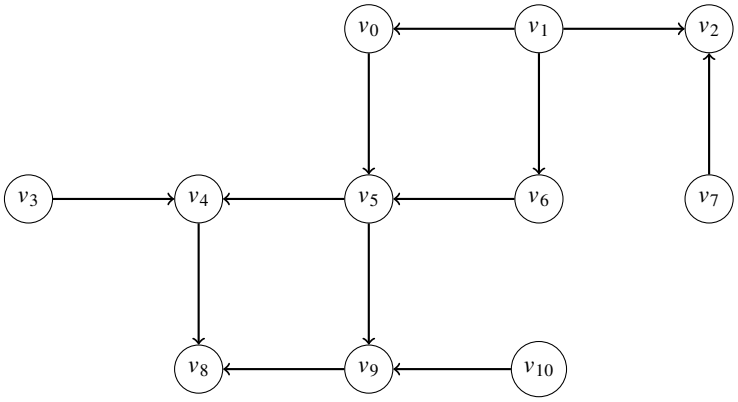


Figura 4.1: Grafo G de entrada para o problema.

Teorema 1 *O problema GIBM não é submodular sobre o modelo MCICM.*

Prova: Dado o grafo $G(V, E)$ da Figura 4.1, que representa o grafo de entrada do problema e N_0 o conjunto das sementes negativas, de modo que $N_0 = \{v_1\}$, é criado um grafo $G^P(V, E^P)$ sendo E^P as arestas ativadas previamente que passam a informação positiva antes do início da simulação, suponha que esse grafo é o da Figura 4.2, ou seja, apenas as arestas (v_3,v_4) e (v_{10},v_9) são ativadas. Além disso, é criado um outro grafo $G^N(V, E^N)$ que contém apenas as arestas ativadas previamente que transmitem a informação negativa. De modo que E^N é o conjunto de arestas ativadas previamente que passam a informação negativa, representado pela Figura 4.3. Nesse caso a prova é por contraexemplo.

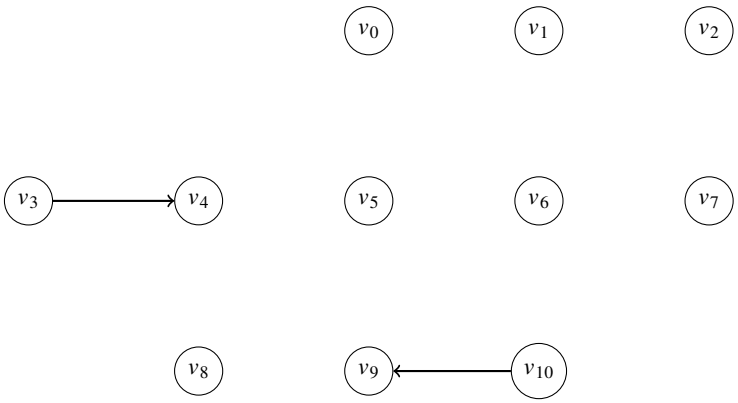


Figura 4.2: Grafo G^P com as arestas ativas que podem passar a informação positiva. Nesse caso usando o conceito de lançamento antes da simulação, temos que apenas duas arestas são capazes de passar a informação positiva.

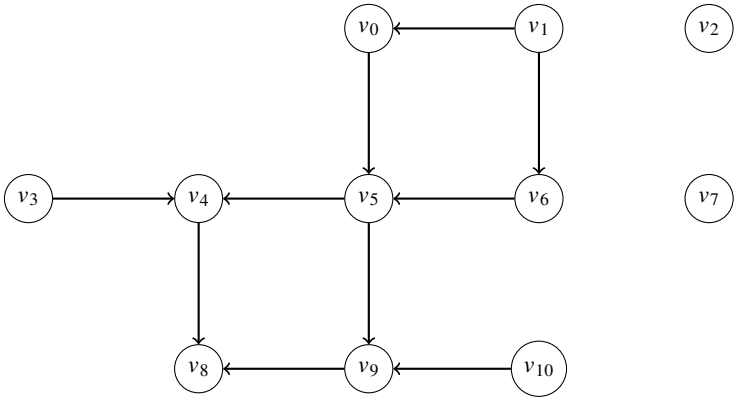


Figura 4.3: Grafo G^N com as arestas ativas que podem passar a informação negativa, traçando assim o caminho da disseminação negativa. É possível observar que o conjunto N_0 consegue alcançar grande parte dos vértices.

Dado que $N_0 = \{v_1\}$, é fácil verificar que no grafo G^P que a melhor solução de escolha de sementes positivas são os vértices v_3 e v_{10}, visto que os outros vértices não possuem arestas de saída. Se escolhermos o vértice v_3, temos $f(\{v_3\}) = 1$, pois o único vértice salvo é v_4, escolhendo o vértice v_{10} como semente, $f(\{v_{10}\}) = 1$, visto que apenas o vértice v_9 é salvo da contaminação. Por outro lado, se escolhermos v_3 e v_{10} como semente ao mesmo tempo, temos que $f(\{v_3, v_{10}\}) = 3$. Pois, as arestas (v_4,v_8) e (v_9, v_8) não são arestas ativas para passar a informação positiva, assim a informação positiva nunca chegará no vértice v_8. Entretanto, o caminho para a informação negativa é bloqueado até o vértice v_8, sendo assim os vértices $\{v_4, v_8, v_9\}$ salvos e isso não é submodular (observe a Figura 4.4), como é demostrado abaixo. Considere $S = \emptyset$, $T = \{v_3\}$ e $v = v_{10}$. Então temos pela equação da submodularidade.

$$f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)$$
$$f(\emptyset \cup \{v_{10}\}) - f(\emptyset) \geq f(\{v_3\} \cup \{v_{10}\}) - f(\{v_3\})$$
$$1 - 0 \geq 3 - 1$$
$$1 \geq 2$$

que é uma contradição. □

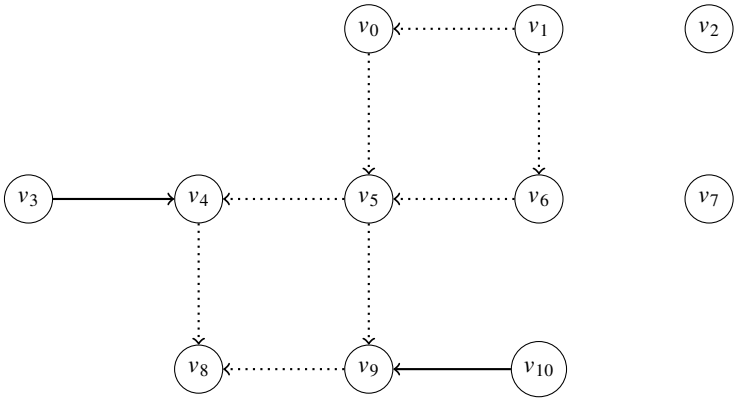


Figura 4.4: Considere as linhas pontilhadas como as arestas que passam a informação negativa e as contínuas que representam o caminho da informação positiva. Na imagem apresentada fica mais claro observar que selecionando os vértices v_3 e v_{10} ao mesmo tempo acaba bloqueando o caminho até o vértice v_8, sendo assim a disseminação negativa não possui um caminho livre até o vértice v_8.

Conforme mostra o Teorema 1, a propriedade de submodularidade não é válida para o modelo MCICM, e portanto o algoritmo guloso que garante aproximação de $1 - 1/e$ não é passível de ser utilizado.

Também é possível abordar o problema que propomos de uma maneira diferente, todos os algoritmos gulosos relatados até aqui não levam em consideração o custo da escolha do vértice na função de ganho de influência. Dado isso é possível trabalhar com um algoritmo para o problema *maximização da influência com orçamento* (em inglês *budgeted influence maximization*), proposto por Nguyen e Zheng (2013) [24], e pode ser entendido como a versão com orçamento do problema *maximização de influência* proposto por Kempe et al. (2003) [17]. O problema consiste em dado um grafo direcionado $G(V, E)$, uma função de custo $C : V(G) \rightarrow Z^+$ e um orçamento fixo $b \in Z^+$. A função de custo C atribui um custo de seleção não uniforme a cada vértice da rede, que é o custo a ser pago para a escolha desse vértice como um vértice de início da disseminação no grafo. O objetivo é selecionar um conjunto de vértices dentro do orçamento, que maximiza a propagação da influência na rede. Mais formalmente, deve-se encontrar um conjunto S tal que $\sum_{u \in S} C(u) \leq b$ de modo que para qualquer conjunto S' com $\sum_{u \in S'} C(u) \leq b$, a seguinte condição seja verdadeira $|\sigma(S)| \geq |\sigma(S')|$.

Algoritmo 3: Algoritmo Guloso Ingênuo

Entrada: Dado um grafo $G(V, E)$ e o orçamento para escolha do conjunto S denotado por b

Saída: Um conjunto de sementes S

```
1  início
2      S = ∅
3      repita
4          δ(v) = (σ(S ∪ v) − σ(S))/c(v), ∀v ∈ V
5          u = argmax_{v∈V} δ(v)
6          se c(S ∪ u) ≤ b então
7              S = S ∪ u
8          fim
9          V = V \ u
10     até V = ∅;
11 fim
12 retorna S
```

Os autores propuseram o Algoritmo 3 seguindo o conceito de adicionar o custo de cada vértice em relação ao seu benefício. A cada rodada o algoritmo adiciona o vértice de maior custo benefício ao conjunto S, de modo que esteja dentro do orçamento b. Além disso, os autores observam que o Algoritmo 3 pode ter razão de aproximação ilimitada. Considere um grafo contendo $l + 1$ vértices, sendo que $V = \{u, v_1, v_2, ..., v_l\}$. Cada par em v_1, v_2,..., v_l está conectado por uma aresta com probabilidade igual a 1, enquanto u é um vértice isolado. Seja o custo $c(u) = 1 - \epsilon$, $c(v_i) = l, \forall i = 1, ..., l$ e o orçamento b para o problema igual a l. A solução ótima irá escolher qualquer vértice v_i e ter como resultado l atingidos. Em contraste, o Algoritmo 3 escolhe u, pois tem sua média de influência de custo igual a $\frac{1}{1-\epsilon} > 1$, pois

$$\frac{\text{total de influência de } u}{\text{total de custo de } u} > \frac{\text{total de influência } c(v_i)}{\text{total de custo } c(v_i)}$$
$$\frac{1}{1-\epsilon} > \frac{l}{l}$$
$$\frac{1}{1-\epsilon} > 1$$

e como a disseminação de influência resultante é igual a 1. Então, a razão de aproximação para Algoritmo 3 é l.

Algoritmo 4: Algoritmo Guloso Melhorado

Entrada: Dado um grafo $G(V, E)$ e o orçamento para escolha do conjunto S denotado por b
Saída: Um conjunto de sementes S

```
início
    S_1 = resultado do algoritmo 3
    S_max = argmax_{v∈V} δ(v)
    S = argmax(σ(S_1), σ(S_max))
fim
retorna S
```

Tentando contornar as deficiências que o Algoritmo 3 produz, os autores apresentam o Algoritmo 4, que escolhe entre o conjunto S resultante do Algoritmo 3 e o vértice que maximiza o alcance na rede sem levar em consideração o custo, escolhendo o que dá maior ganho de alcance na rede. Segundo os autores, o Algoritmo 4 garante uma aproximação de $(1 - 1/\sqrt{e})$ para o problema *maximização da influência com orçamento*. A seguir, no Teorema 2, vamos mostrar que o GIBM também pode se beneficiar de tal fato.

Teorema 2 *Maximização do bloqueio da influência generalizado apresenta uma aproximação de* $(1 - 1/\sqrt{e})$ *sobre o modelo COICM com o Algoritmo 4.*

Prova: No modelo COICM tanto a disseminação positiva como a negativa tem a mesma probabilidade de se espalhar pela rede. Nesse caso, cada aresta tem um único peso $w_{u,v}$ que indica a probabilidade da disseminação avançar pela rede. Seguindo a mesma abordagem de Kempe et al. (2003) [17] que foi usada na demonstração do Teorema 1, podemos pré-lançar todas as moedas para determinar quais arestas do grafo G estão ativas e traçar o caminho das disseminações. A prova é similar à de Budak et al. (2011) [5] para o problema *limitação eventual de influência*.

Dado um grafo $G(V, E)$ da Figura 4.5, que representa o grafo de entrada do problema e N_0 o conjunto das sementes negativas, o primeiro passo é criar um grafo $G'(V, E')$ onde E' é o conjunto das arestas ativadas, definidas previamente com lançamento das moedas antes do início da simulação. Temos que tanto a disseminação positiva como a negativa podem ser modeladas no grafo G', pois a ativação de uma aresta nesse modelo indica que ela pode passar a informação positiva ou negativa, o que vier primeiro. Como o grafo G' pode simular a disseminação negativa é possível saber quais vértices são alcançáveis a partir da origem da disseminação negativa, chamamos o conjunto desses vértices que são alcançáveis a partir de N_0 como N'. Seja esse grafo representado pela Figura 4.6.

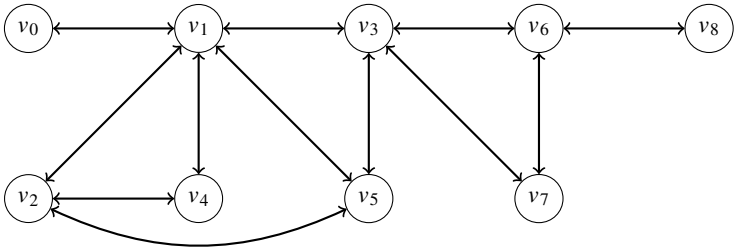


Figura 4.5: Considere o grafo G da imagem como o de entrada para o GIBM.

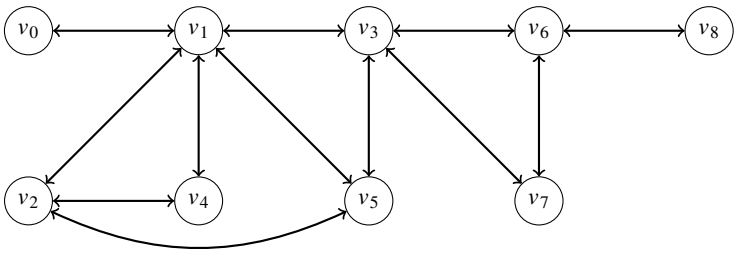


Figura 4.6: Considere que o conjunto que da ao início da disseminação negativa seja igual a $N_0 = \{v_0, v_5\}$. Ao lançar todas as moedas para verificar qual aresta fica ativa para passar informações suponha que todas as arestas do grafo sejam ativadas. Logo, o grafo G' é similar ao grafo da Figura 4.5 e $N' = \{v_1, v_2, v_3, v_4, v_6, v_7, v_8\}$.

A seguir, criamos um grafo G'' que representa onde a informação positiva chega antes da informação negativa. Seja $P(u, v)$ o conjunto de arestas dos caminhos mais curtos de u a v em G'. Assim como, $P(N_0, v)$ o conjunto de arestas dos caminhos mais curtos do vértice mais próximo em N_0 a v em G'. Formalmente, $P(N_0, v) = \{P(u, v) : u = \text{argmin}_{u' \in N_0} |P(u', v)|\}$. Então, $G''(N'', E'')$ é definido como

$$N'' = V \setminus N_0$$
$$E'' = \{P(u, v) \ : \ |P(u, v)| < |P(N_0, v)| \ \ \forall\, v \in N', u \in N''\}$$

Intuitivamente, G'' adiciona os caminhos mais curtos de u a v em G' se tais caminhos chegarem antes de um vértice de N_0 a v. As Tabelas 4.1, 4.2 e 4.3 ilustram o processo de construção passo a passo de E'' baseado no exemplo da Figura 4.5. Por fim, a Figura 4.7 mostra o grafo final G'' para este exemplo.

N'' \ N'	v_1	v_2	v_3	v_4	v_6	v_7	v_8
v_1	X	1	1	1	2	2	3
v_2	1	X	2	1	3	3	4
v_3	1	2	X	2	1	1	2
v_4	1	1	2	X	3	3	4
v_6	2	3	1	3	X	1	1
v_7	2	3	1	3	1	X	2
v_8	3	4	2	4	1	2	X

Tabela 4.1: Para montar o conjunto de arestas do grafo G'' o processo de escolha é mais perceptível quando é construindo a matriz de distâncias entre os conjuntos N' e N'' no grafo G'. Sabemos que $N'' = \{v_1, v_2, v_3, v_4, v_6, v_7, v_8\}$, pois $N'' = \{u | u \in V \setminus N_0\}$.

N_0 \ N'	v_1	v_2	v_3	v_4	v_6	v_7	v_8
v_0	1	2	2	2	3	3	4
v_5	1	1	1	2	2	2	3

Tabela 4.2: Tabela que contém as distâncias dos vértices pertencentes ao conjunto N_0 até os vértices do conjunto N'

A ideia é que o grafo G'' é o grafo em que a disseminação positiva chega antes da negativa. Agora, está claro que resolver o problema GIBM é equivalente a maximizar o número de vértices alcançáveis de um conjunto inicial S em G'', mas o último é precisamente o problema *maximização da influência com orçamento*. □

$v \in N'$	$\lvert P(N_0, v)\rvert$	Caminho adicionado em G'' (Se $\lvert P(u,v)\rvert < \lvert P(N_0,v)\rvert, \forall u \in N''$)
v_1	1	X
v_2	1	X
v_3	1	X
v_4	2	$v_1 \rightarrow v_4, v_2 \rightarrow v_4$
v_6	2	$v_3 \rightarrow v_6, v_7 \rightarrow v_6, v_8 \rightarrow v_6$
v_7	2	$v_3 \rightarrow v_7, v_6 \rightarrow v_7$
v_8	3	$v_3 \rightarrow v_8, v_6 \rightarrow v_8, v_7 \rightarrow v_8$

Tabela 4.3: Esta tabela mostra as arestas adicionadas em G'', com base nos valores das Tabelas 4.1 e 4.2. Por exemplo, para o vértice v_4, o caminho mais curto de N_0 para ele é 2 (ou seja, o valor $\lvert P(N_0, v_4)\rvert$, recuperado da Tabela 4.2), tal que todos os caminhos mais curtos $P(u, v)$ com distância menor que 2 (tais distâncias são recuperadas da Tabela 4.1) são adicionados em E''. Neste caso, $v_1 \rightarrow v_4$ e $v_2 \rightarrow v_4$. Isso significa que, ao escolher v_1 ou v_2 como sementes positivas, a informação positiva chega antes da negativa em v_4.

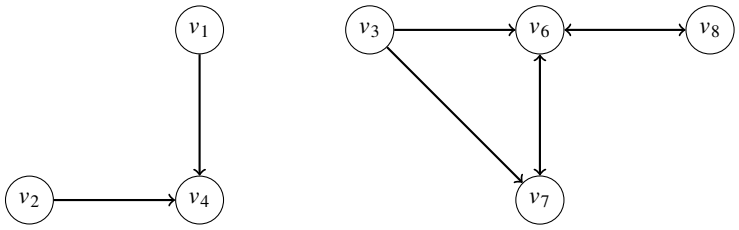


Figura 4.7: Grafo G'', note que nesse grafo queremos maximizar o alcance, pois é o grafo onde a disseminação da informação positiva chega antes da informação negativa. Logo queremos escolher o melhor conjunto que esteja dentro de um orçamento que aumente o alcance nesse grafo e esse é o problema de *maximização de influência com orçamento*.

Com isso, a partir de um grafo de entrada para o GIBM é possível manipulá-lo de maneira que esse grafo seja uma entrada para o problema *maximização da influência com orçamento*, sendo possível executar o Algoritmo 4 e verificar qual seria o melhor conjunto de vértices para iniciar a disseminação positiva. Por fim, é possível estender a prova de He et al. (2011) [16] para o GIBM quando tratado sobre o modelo de limitante linear competitivo. Nesse caso *maximização do bloqueio da influência* é o caso uniforme do problema que abordamos. Com isso, é possível afirmar que o problema GIBM é submodular também, desde que o orçamento da entrada do problema permita escolher os k vértices que maximizam a diminuição da disseminação da informação negativa, visto que na demonstração da submodularidade não é levada em consideração o orçamento.

Teorema 3 *O problema GIBM garante a aproximação de* $1 - \frac{1}{e}$ *no modelo limitante linear competitivo.*

Prova: A prova do Teorema 3 é omitida por ser idêntica àquela provada por He et al. (2011) [16]. □

Desse modo, foi possível concluir três propriedades para o *maximização do bloqueio de influência generalizado*, foi possível chegar a conclusão que para o modelo de cascata MCICM o problema não apresenta submodularidade. Quando tratado o modelo de COICM concluimos que é equivalente ao problema *maximização da influência com orçamento*. Ademais foi demostrado que para o modelo de limitante linear competitivo a propriedade de submodularidade é válida, garantindo o fator de aproximação. No próximo capítulo iremos focar em experimentos usando medidas de centralidade para a escolha do conjunto de vértices que dão início a disseminação positiva.

5 EXPERIMENTOS E RESULTADOS

Neste capítulo explicaremos com detalhes a metodologia usada e um detalhamento mais profundo sobre as medidas de centralidades adotadas. Ademais, serão discutidos os resultados usando tais medidas para o problema *maximização do bloqueio de influência generalizado* com duas funções de custo diferentes.

5.1 METODOLOGIA

Os estudos neste trabalho têm como um dos objetivos investigar várias medidas de centralidade a serem usadas como estratégias para resolver o problema nos casos das funções de custo *uniforme* e *penalização do grau* no modelo de disseminação MCICM. No primeiro caso que é da função de custo *uniforme*, imitamos um caso (um tanto irreal) que já foi considerado em trabalhos anteriores, enquanto no segundo buscamos um cenário mais realista, onde o custo de um vértice é proporcional ao seu grau.

Para ambos os casos de funções de custo, realizamos simulações em grafos direcionados e não direcionados, com o objetivo de analisar o comportamento entre os dois tipos de grafos. Se o grafo de entrada não estiver totalmente conectado, levaremos em consideração apenas a maior componente conectada (ou a maior componente fracamente conexa para grafos direcionados) do grafo. Essa é uma prática comum, pois as informações corretas e as informações incorretas não podem pular de uma componente para outra. Nas simulações, analisamos três cenários diferentes para a probabilidade de uma informação/desinformação ser propagada. Mais precisamente, realizamos experimentos para as probabilidades w^+ e w^- da seguinte maneira:

- **Propagação baixa:** w^+ e w^- são escolhidos no intervalo $[0, 0.2]$
- **Propagação normal:** w^+ e w^- são escolhidos no intervalo $[0, 1]$.
- **Propagação alta:** w^+ e w^- são escolhidos no intervalo $[0.75, 1]$.

Nos três casos, os valores w^+ e w^- são escolhidos de forma independente e aleatória com uma distribuição uniforme. Os vértices do conjunto de sementes negativas N_0 são escolhidos de maneira uniformemente aleatória no grafo, sendo que fixamos o tamanho do conjunto N_0 igual a 1% da soma da função de custo de todos os vértices e variamos o tamanho do conjunto S em 0.1%, 0.5%, 1%, 1.5% e 2.0% em relação a soma total do custo dos vértices de cada grafo.

Para os experimentos escolhemos três conjuntos de dados do mundo real, entre os quais duas são redes de citações (DBLP e CORA) e um conjunto de dados que representa uma eleição interna da plataforma Wikipédia. A rede de citação DBLP é um conjunto de dados de publicações científicas, como artigos e livros [22], em que um vértice representa uma publicação e uma aresta representa uma citação, ou seja, existe uma aresta de A a B se o artigo A citar a publicação B. O banco de dados DBLP original possui 12.590 vértices e 49.749 arestas. O conjunto de dados Cora contém mais de 23.000 vértices e aproximadamente 90.000 arestas [29]. Semelhante ao DBLP, esse conjunto de dados representa citações em artigos de uma plataforma, em que os vértices são artigos e as arestas são citações entre eles. O conjunto de dados de eleições da Wikipédia representa a enciclopédia colaborativa Wikipédia em inglês, de usuários que votaram a favor ou contra o outro nas eleições de administrador, vértices representam usuários e uma aresta representa um usuário que votou em outro [19]. Todos os conjuntos de dados descrevem originalmente grafos direcionados.

Os mesmos conjuntos de dados foram usados nos experimentos em grafos não direcionados, com a direção das arestas sendo ignorada. A opção de ignorar a direção das arestas, em vez de usar outros conjuntos de dados originalmente não direcionados, permite uma comparação mais direta entre os casos direcionados e não direcionados. A Tabela 5.1 mostra a comparação dos três conjuntos de dados após o pré-processamento para encontrar a maior componente.

Rede	Vértices	Arestas
CORA	23,166	89,157
DBLP	12,495	49,578
Wikipédia Eleição	7,066	100,721

Tabela 5.1: Visão Geral dos Grafos.

Para cada cenário proposto, usamos as seguintes medidas de centralidade como uma contra-estratégia:

- Coeficiente de Clustering;
- PageRank;
- Centralidade de Intermediação;
- Centralidade de Percolação.

Além das medidas acima, usamos duas estratégias para o controle dos experimentos: escolhendo vértices de maior grau primeiro (de maneira gulosa) e escolhendo vértices de maneira aleatória. A seguir as medidas de centralidade citadas acima serão detalhadas com mais profundidade.

5.1.1 Coeficiente de Clustering

Em teoria dos grafos, o coeficiente de clustering [12] avalia o grau com que os vértices de um grafo tendem a se agrupar. Em uma rede social online como a Internet, por exemplo, os indivíduos são capazes de se conectar com facilidade, independentemente de suas diferenças culturais, sociais ou distância geográfica. Entretanto, as associações não tendem a ocorrer de forma aleatória, mas sim entre pessoas com afinidades semelhantes. Estas afinidades possibilitam a geração de clusters. Em grafos sem pesos o coeficiente de clustering de cada vértice v é a fração de triângulos que o vértice v participa pela quantidade de triângulos possíveis, isto é:

$$C_v = \frac{2T(v)}{deg(v)(deg(v)-1)}$$

onde $T(v)$ é a quantidade de triângulos que o vértice v faz parte e $deg(v)$ é o grau de v.

5.1.2 PageRank

O PageRank [25] calcula uma classificação para os vértices com base na estrutura dos *links* recebidos, ou seja, arestas de entrada. Foi originalmente projetado como um algoritmo para classificar páginas da web. Neste caso, cada vértice é visto como uma página web e as arestas os *links* que direcionam uma a outra.

A medida PageRank pode ser entendida com um fluido que percorre pelos vértices através das arestas e que acaba se acumulando nos vértices mais importantes da rede. O cálculo para cada vértice é feito da seguinte maneira.

- Dado um número k, que representa a quantidade de atualizações e um grafo com n vértices, temos que inicialmente todos os vértices começam com um PageRank de $\frac{1}{n}$
- Cada etapa k, consiste em cada vértice dividir igualmente seu PageRank com suas arestas de saída, assim transmitindo o mesmo valor para todas elas. Após isso cada vértice atualiza seu PageRank somando a quantidade que recebeu.
- O processo termina após k etapas, sendo os vértices com maior PageRank da rede os mais importantes.

5.1.3 Centralidade de Intermediação

A intermediação é uma medida de centralidade (em inglês *betweenness*) que leva em consideração que um vértice importante da rede deve fazer parte de vários caminhos mínimos entre pares de vértices de um grafo [4]. Dado um grafo $G = (V, E)$, onde V representa o conjunto dos vértices e E o conjunto de arestas do grafo G, temos que a centralidade de intermediação de um vértice v é a soma da fração dos caminhos mínimos de todos os pares de vértices que passam por v, ou seja

$$C_B(v) = \sum_{s,t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

onde σ_{st} é o número de caminhos mínimos de s até t, e $\sigma_{st}(v)$ é o número destes caminhos que passam pelo vértice v, dado que v é diferente de s ou t. Além disso temos que se $s = t$, $\sigma_{s,t} = 1$, e se $v \in \{s, t\}$, então $\sigma_{st}(v) = 0$.

5.1.4 Centralidade de Percolação

O conceito de percolação ocorre em redes complexas em uma série de cenários. Por exemplo, a infecção viral ou bacteriana pode se espalhar pelas redes sociais de pessoas, boatos ou notícias sobre ofertas de negócios também podem se espalhar pelas redes sociais de pessoas. Em todos esses cenários, um "contágio" se espalha pelas arestas de uma rede complexa, alterando os "estados" dos vértices à medida que se espalha. Por exemplo, em um cenário epidemiológico, os indivíduos vão de um estado "suscetível" a "infectado" conforme a infecção se espalha, como aborda Newman (2010) [23]. A característica comum em todos esses cenários é que a propagação do contágio resulta na mudança dos estados dos vértices nas redes. Com isso é possível definir a centralidade de percolação proposta por Piraveenan et al. (2013) [27] que quantifica o impacto relativo dos vértices com base em sua conectividade topológica, bem como em seus estados de percolação.

Essa medida generaliza a centralidade de intermediação atribuindo pesos de percolação para cada vértice do grafo. Então dado um grafo $G = (V, E)$, o vértice $v \in V$, os estados de percolação X_v, para todo $v \in V$ e $R(x) = max\{x, 0\}$. Seja σ_{uw} a quantidade de caminhos mínimos de u a w e $\sigma_{uw}(v)$ a quantidade de caminhos mínimos de u a w que passam por v. Então a centralidade de percolação do vértice v é o seguinte

$$\text{perc}(v) = \sum_{\substack{(u,w) \in V^2 \\ u \neq v \neq w}} \frac{\sigma_{uw}(v)}{\sigma_{uw}} \frac{R(x_u - x_w)}{\sum_{\substack{(f,d) \in V^2 \\ u \neq v \neq w}} R(x_f - x_d)}$$

Como a centralidade de percolação requer estados de percolação para os vértices que refletem um certo grau de "contaminação", usamos uma medida em nossos experimentos para

definir os pesos de percolação para cada vértice da seguinte maneira. Seja $d(v, N_0)$ a distância de v ao vértice mais próximo em N_0. Assim, o peso da percolação x_v para o vértice v é definido como

$$x_v = \frac{1}{d(v, N_0) + 1}$$

A ideia é que os vértices inicialmente em N_0 sejam 100% percolados (neste caso, $d(v, N_0) = 0$) e quanto mais longe de N_0 um vértice esteja, seu peso de percolação diminui. Como exemplo do cálculo dos pesos de percolação para cada vértice consideramos o grafo da Figura 5.1 e seja $N_0 = \{v_0, v_4\}$, temos que o peso de cada percolação será igual ao mostrado na Figura 5.2.

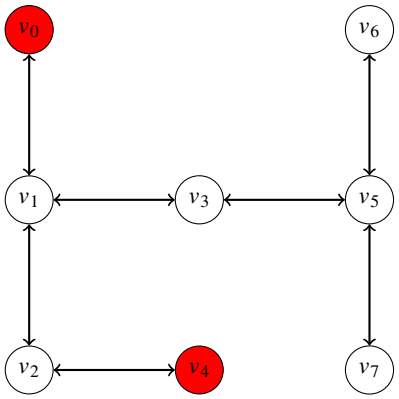


Figura 5.1: Grafo G usado no exemplo para cálculo dos pesos de percolação.

Vértice	Peso
v_0	1
v_1	0,50
v_2	0,50
v_3	0,33
v_4	1
v_5	0,25
v_6	0,20
v_7	0,20

Figura 5.2: Pesos de percolação para cada vértice do grafo da Figura 5.1

5.1.5 Informações Complementares

Os experimentos foram executados em uma CPU Intel (R) Core (i) i7-6700 a 3,40 GHz e 8 GB de RAM. Os scripts foram implementados na linguagem Python 3.6.9. Para manipulações de grafos, foi usada a biblioteca NetworkX 2.3 [15]. A implementação de todas as medidas de rede consideradas neste trabalho também está disponível na documentação da NetworkX, assim como as referências de implementação para cada uma.

5.2 RESULTADOS

Nesta seção, avaliamos o desempenho de diferentes estratégias para encontrar uma solução para o problema GIBM. Como o MCICM é um modelo probabilístico, realizamos experimentos repetidos para espalhar os conjuntos iniciais N_0 e S e obter o comportamento médio. Em cada cenário diferente, realizamos a simulação 1000 vezes para obter a média dos conjuntos contaminados positivamente e negativamente.

Inicialmente comparamos quatro cenários diferentes, como mostra a Tabela 5.2. Lembramos que para cada cenário há três casos para a probabilidade de disseminação.

Função de Custo	Tipo de Grafo	Probabilidade de Propagação
uniforme	Não Direcionado	[0,0.2], [0, 1] e [0.75,1]
	Direcionado	[0,0.2], [0, 1] e [0.75,1]
penalização de grau	Não Direcionado	[0,0.2], [0, 1] e [0.75,1]
	Direcionado	[0,0.2], [0, 1] e [0.75,1]

Tabela 5.2: Cenários dos experimentos.

5.2.1 Custo Uniforme

Nesta seção, mostramos e analisamos os resultados obtidos para a função de custo *uniforme*. Para esses experimentos, definimos o tamanho de N_0 como 1% do número de vértices de cada conjunto de dados e variamos o parâmetro k (aqui o tamanho da saída definida como S para sementes positivas é igual a k) entre 0,1%, 0,5%, 1%, 1,5% e 2,0% do número de vértices em cada conjunto de dados. Analisamos os casos para os grafos não direcionados e direcionados. Em cada grafo, mostramos as médias dos resultados para os três grafos. No eixo vertical, mostramos a porcentagem de vértices contaminados negativamente, portanto, quanto menores os valores, melhor a métrica funciona como uma estratégia para o problema. O número absoluto de vértices em função da porcentagem pode ser visto na Tabela 5.3.

Network	**0.1%**	**0.5%**	**1%**	**1.5%**	**2%**
CORA	23	115	231	347	463
DBLP	12	62	124	187	249
Wiki	7	35	70	105	141

Tabela 5.3: Tamanho de k (em função da % de $|V|$).

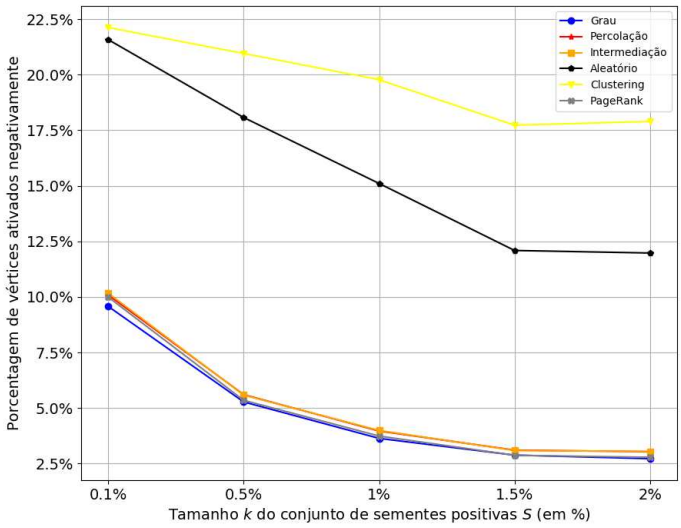


Figura 5.3: Função de custo *uniforme* em grafos não direcionados com baixa propagação.

Nas Figuras 5.3, 5.4 e 5.5 apresentamos os resultados para a função de custo *uniforme* em grafos não direcionados com baixa, normal e alta probabilidade de propagação, respectivamente.

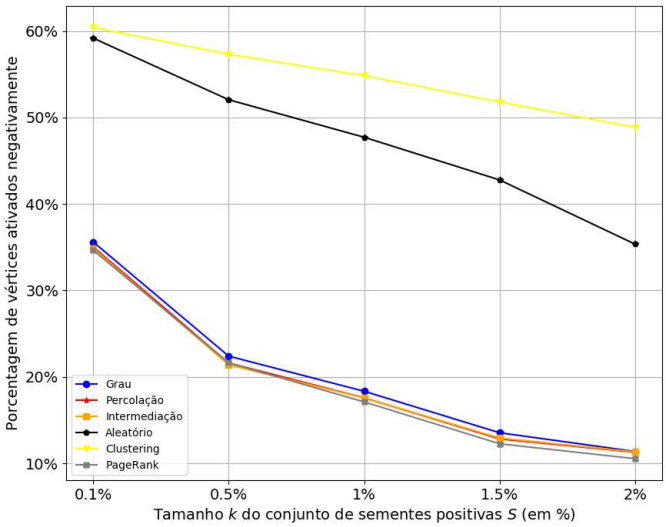


Figura 5.4: Função de custo *uniforme* em grafos não direcionados com média propagação.

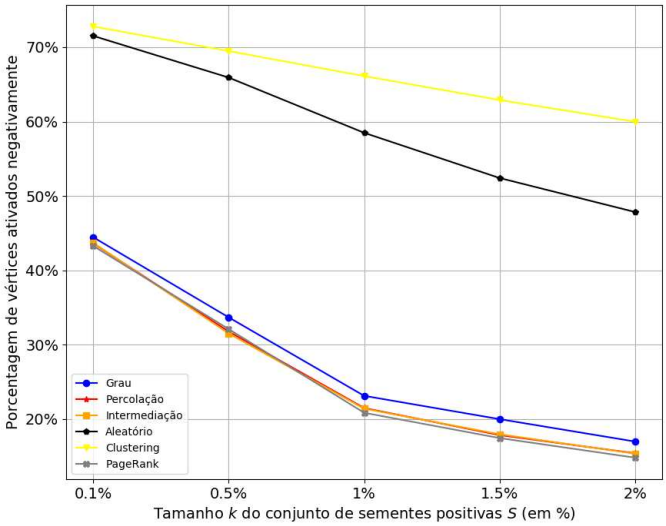


Figura 5.5: Função de custo *uniforme* em grafos não direcionados com alta propagação.

No caso de baixa probabilidade de propagação (Figura 5.3), vemos que as medidas de percolação, intermediação, grau e PageRank se comportam de maneira semelhante e também apresentam desempenho melhor do que outras estratégias. No cenário de probabilidade de propagação normal (Figura 5.4), temos uma maior propagação negativa em comparação com o caso de baixa probabilidade (o que é esperado, uma vez que a probabilidade de propagação é maior), mas a qualidade das estratégias permanece consistente com o caso anterior. No caso em que a rede é altamente influenciável (Figura 5.5), notamos que a centralidade do grau tem um desempenho diminuído em comparação aos casos anteriores e com as medidas de intermediação, percolação e PageRank. Além disso, as estratégias de clustering e aleatória apresentam resultados pífios nos três casos.

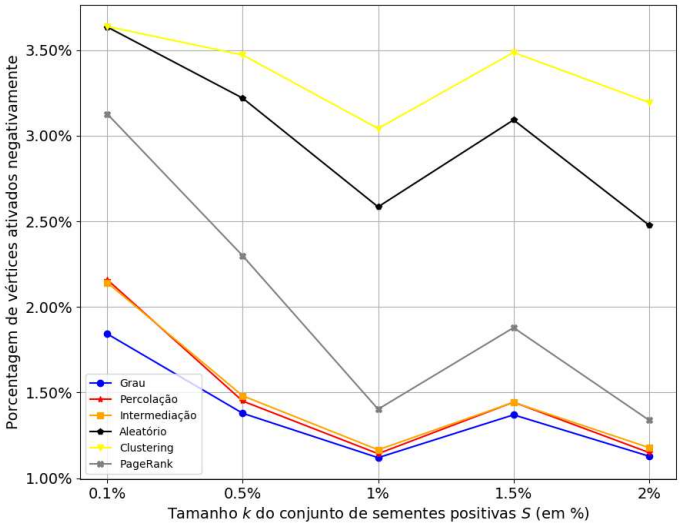


Figura 5.6: Função de custo *uniforme* em grafos direcionados com propagação baixa.

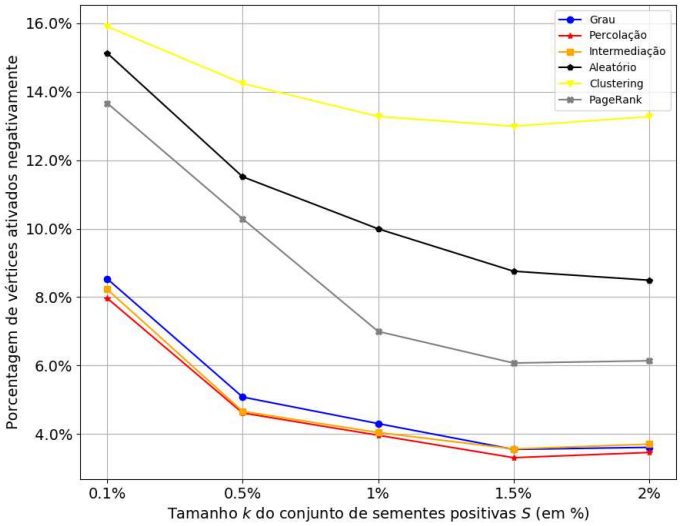


Figura 5.7: Função de custo *uniforme* em grafos direcionados com propagação normal.

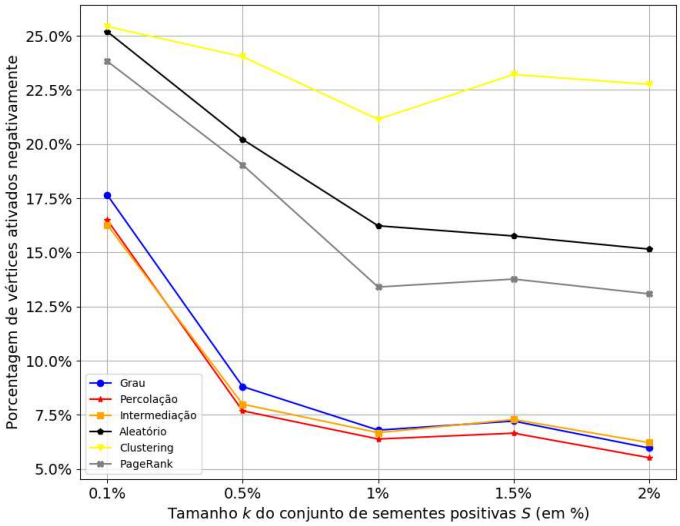


Figura 5.8: Função de custo *uniforme* em grafos direcionados com propagação alta.

No caso de grafos direcionados, o número de vértices com influência positiva diminui em comparação com a versão não direcionada, conforme mostrado nas Figuras 5.6, 5.7 e 5.8. No cenário de baixa probabilidade (Figura 5.6), há uma ligeira melhora na qualidade da centralidade do grau, no entanto, podemos considerar que as estratégias de percolação, grau, intermediação e PageRank novamente tem desempenhos estatisticamente equivalentes dentro de uma pequena margem de erro. Com probabilidade normal de propagação (Figura 5.7), o comportamento do PageRank aparentemente piora em comparação com os casos anteriores. O último caso para grafos direcionados na função de custo *uniforme* (Figura 5.8) é semelhante ao caso anterior (probabilidade normal), com a única diferença sendo o maior número de vértices influenciados.

A hipótese é que os comportamentos semelhantes entre grau, intermediação, PageRank e percolação provêm do fato de que o conjunto de sementes positivas escolhidas por essas estratégias é semelhante. Para testar essa hipótese, tomamos o conjunto de sementes positivas da centralidade do grau como base para a comparação e medimos a semelhança entre os conjuntos retornados pelas outras estratégias. Mais formalmente, sejam S_1 e S_2 os conjuntos retornados usando, respectivamente, a centralidade do grau e alguma outra estratégia. Para medir a semelhança entre os conjuntos, usamos o *coeficiente de sobreposição* em inglês definido como *overlap coefficient*, definido como

$$\text{overlap}(S_1, S_2) = \frac{|S_1 \cap S_2|}{\min\{|S_1|, |S_2|\}}.$$

A Figura 5.9 mostra os resultados das semelhanças entre as soluções no conjunto de dados DBLP. No eixo vertical, temos o coeficiente de sobreposição, tomando a centralidade do grau como comparação de base. O eixo horizontal representa o tamanho do conjunto, e mostramos soluções de até 250 vértices, pois esse é aproximadamente o tamanho dos conjuntos maiores para as soluções dos experimentos. Observamos que as soluções que utilizam intermediação, percolação e PageRank como estratégias têm um alto coeficiente de *overlap*. Isso significa que os conjuntos de soluções retornados por essas estratégias são semelhantes à estratégia de centralidade do grau. Por outro lado, as soluções obtidas usando o coeficiente de agrupamento

e a amostragem aleatória como estratégia têm um *overlap* muito pequeno, portanto são muito diferentes dos vértices com maiores graus do grafo.

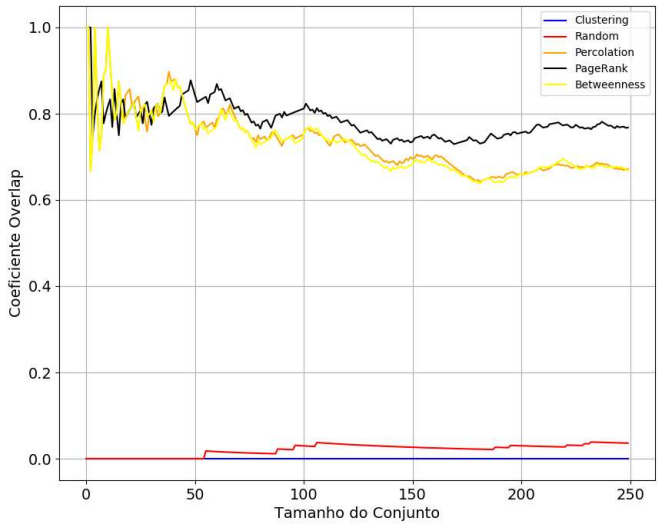


Figura 5.9: Coeficiente *overlap* na base de dados DBLP sem direção: o valor 0 é o caso em que os elementos da solução são completamente diferentes dos vértices de k maiores grau.

5.2.2 Penalização de Grau

Nesta seção, analisamos os resultados da função de custo *penalização de grau*. Nesse caso, os custos são diretamente proporcionais ao grau, portanto, definimos os tamanhos de N_0 e S como uma fração da soma dos graus (ou seja, duas vezes o número de arestas). Mais especificamente, definimos o tamanho de N_0 como igual a 1% da soma dos graus e escolhemos k como 0,1%, 0,5%, 1%, 1,5% e 2% dessa mesma soma. Na Tabela 5.4, mostramos o parâmetro k para cada cenário percentual.

Network	**0.1%**	**0.5%**	**1%**	**1.5%**	**2%**
CORA	178	891	1783	2674	3566
DBLP	99	495	991	1487	1983
Wiki	201	1007	2014	3021	4028

Tabela 5.4: Tamanho de k (em função da % de $2|E|$).

Inicialmente analisamos os resultados para grafos não direcionados. Diferentemente do caso com função de custo *uniforme*, em que o grau de vértice é o atributo central que caracteriza o sucesso de uma determinada estratégia, na função de custo de *penalização de grau*, o peso do vértice "amortiza" a vantagem que o grau exerce nessas estratégias em que vértices de grau alto são priorizados, ou seja, intermediação, percolação e PageRank (lembre-se da Figura 5.9, onde mostramos o coeficiente *overlap* de tais estratégias com o conjunto de vértices de mais alto grau). Portanto, essas estratégias não são tão bem-sucedidas no cenário usando a função de custo *penalização de grau*, como mostrado nas Figuras 5.10, 5.11 e 5.12 (para propagação baixa, normal e alta, respectivamente). De um modo geral, em comparação com a função de custo

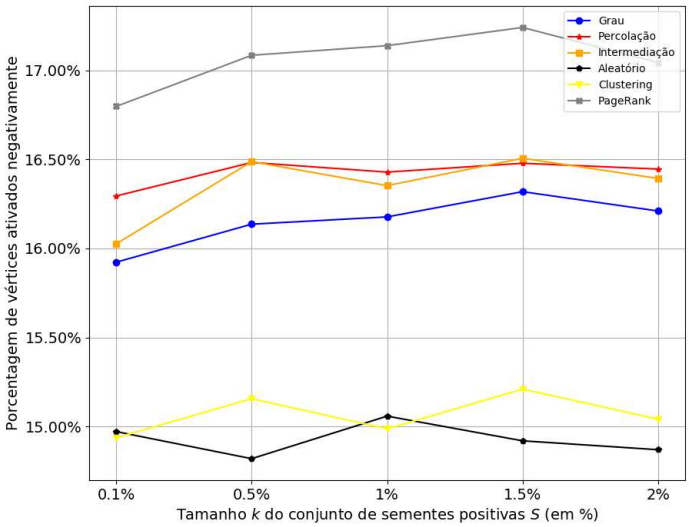


Figura 5.10: Função de custo *penalização de grau* em grafos não direcionados com propagação baixa.

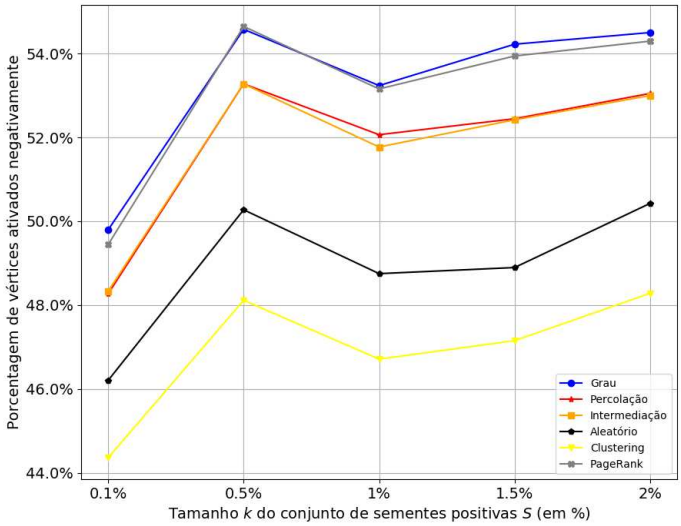


Figura 5.11: Função de custo *penalização de grau* em grafos não direcionados com propagação normal.

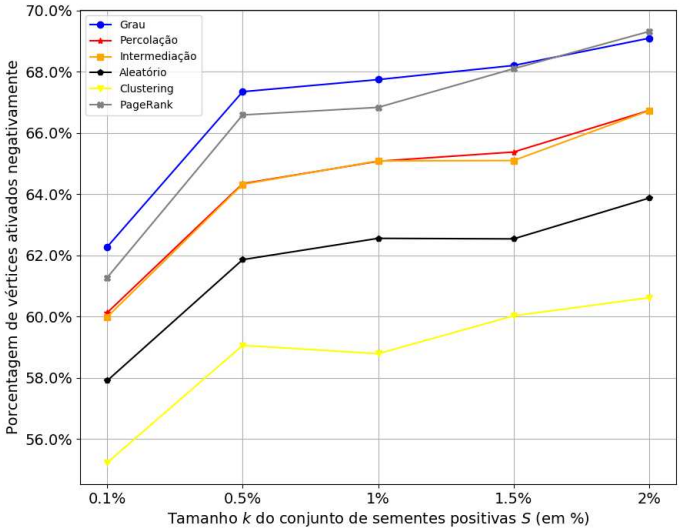


Figura 5.12: Função de custo *penalização de grau* em grafos não direcionados com propagação alta.

uniforme, a função de custo *penalização de grau* influenciou mais negativamente os vértices nas três configurações da probabilidade de disseminação. Além disso, no problema da função de custo *penalização de grau*, a estratégia aleatória e de clustering apresentam os melhores desempenhos entre as métricas que escolhemos.

Em particular, acreditamos que o bom desempenho do coeficiente de clustering pode ser explicado pela diferença entre essa medida e a centralidade do grau (também mostrado na Figura 5.9). Nos conjuntos de dados analisados, os vértices com o maior coeficiente de clustering são aqueles com o menor grau. Como a estratégia que usa o coeficiente de clustering seleciona vértices com baixo grau, isso significa que ele escolhe um grande número de vértices para a solução, pois o custo dos vértices nesse caso é baixo. Portanto, a estratégia do coeficiente de clustering pode ter êxito ao escolher uma fração alta dos vértices de um grafo.

A estratégia aleatória também teve bons resultados e temos alguma suposição para o seu sucesso. Como, em grafos do mundo real, normalmente, a distribuição de graus é aproximada por uma distribuição de lei de potência, e a grosso modo, esses grafos contêm um grande número de vértices de baixo grau. Isso pode explicar, em parte, o bom desempenho da métrica de escolha aleatória. A ideia é que, ao selecionar aleatoriamente os vértices do grafo, a grande maioria são vértices de grau baixo e, portanto, mais vértices são selecionados até atingir o limite máximo do orçamento k.

O comportamento da função de custo *penalização de grau* nos grafos direcionados é diferente dos outros casos analisados até o momento. A Figura 5.13 mostra o caso em que a rede tem uma baixa probabilidade de propagação. Nessa figura, vemos que as métricas têm praticamente o mesmo desempenho, com exceção do PageRank, que apresenta um desempenho ligeiramente inferior. A Figura 5.14 considera a probabilidade de propagação normal. Nesse caso, intermediação e percolação mostram um resultado melhor do que as outras métricas. Finalmente, a Figura 5.15 mostra os resultados em um ambiente altamente influente. As métricas de percolação e intermediação continuam apresentando bons resultados, no entanto, a estratégia aleatória também acaba tendo um resultado próximo a elas.

A estratégia do coeficiente de clustering teve desempenhos opostos nos casos direcionados e não direcionados. Uma possível explicação é que, no caso direcionado, os vértices

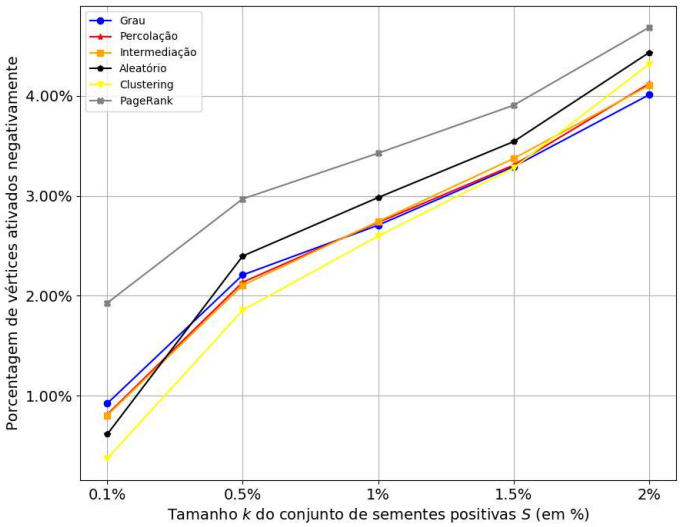


Figura 5.13: Função de custo *penalização de grau* em grafos direcionados com propagação baixa.

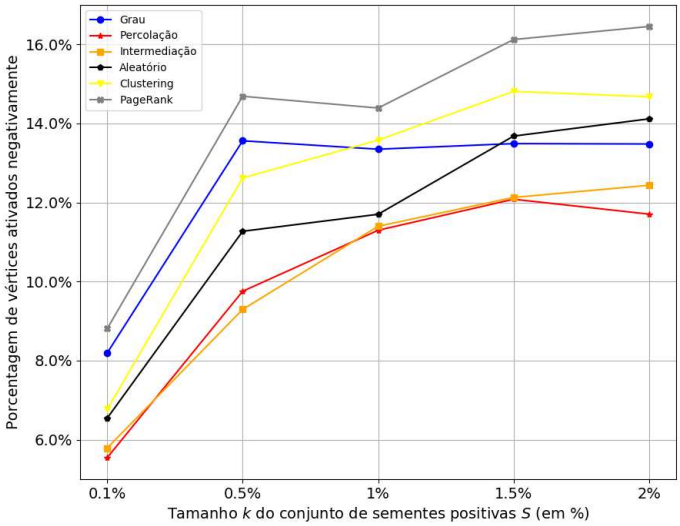


Figura 5.14: Função de custo *penalização de grau* em grafos direcionados com propagação normal.

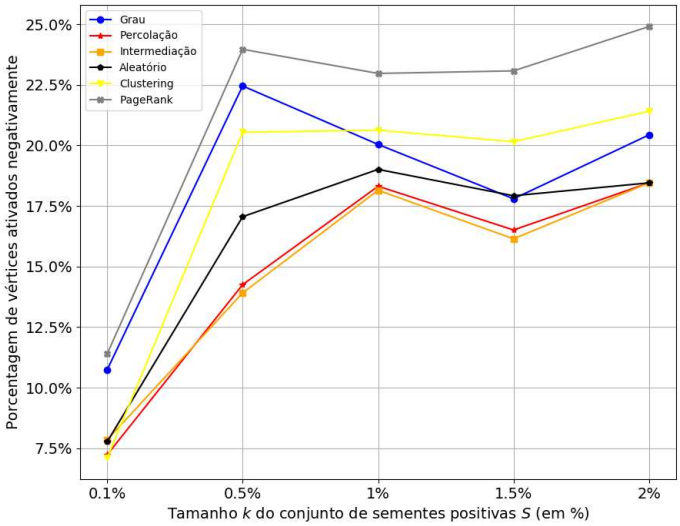


Figura 5.15: Função de custo *penalização de grau* em grafos direcionados com propagação alta.

escolhidos por essa estratégia apresentam um grau baixo. Assim, devido às direções das arestas, muitos podem ter grau de saída igual a zero, impossibilitando a propagação pela rede.

6 CONSIDERAÇÕES FINAIS

O combate a notícias falsas tem se tornado cada vez mais importante devido ao aumento crescente de compartilhamento em redes sociais. Este trabalho forneceu e contextualizou o estado da arte da literatura que buscam minimizar o alcance destas notícias através do contexto de disseminação de informações em uma rede social, especificamente do ponto de vista matemático e computacional.

Em trabalhos anteriores, foram propostos os problemas *maximização do bloqueio de influência* e *limitação eventual de influência*, que forneceram bases para a proposta do problema *maximização do bloqueio de influência generalizado* apresentado nesse trabalho. Para ambos os problemas os autores demostraram propriedades da função de influência, o que garante a aproximação de $1 - \frac{1}{e}$ em relação à solução ótima dos problemas.

Além disso, foi possível tirar importantes conclusões sobre propriedades para o problema que é introduzido neste trabalho, como a demostração que para o modelo MCICM não é possível garantir uma aproximação de $1 - \frac{1}{e}$, pois não é submodular. Entretanto, ao analisar o mesmo problema no modelo de limitante linear competitivo seguindo os mesmos passos de He et al. (2011) [16], foi possível estender que a propriedade é válida, partindo da suposição que o orçamento permite a escolha das sementes positivas. Ademais, foi abordado o que acontece com o problema quando considerado o ganho médio de diminuição da informação negativa na rede para o modelo COICM. Nesse caso foi possível criar um novo grafo através de manipulações, sendo possível usar algoritmos para o problema *maximização de influência com orçamento* para o problema GIBM.

Através de experimentos, analisamos o comportamento de estratégias tendo como base métricas de centralidades de rede conhecidas para duas funções de custo específicas: *uniforme* e *penalização de grau*. Além disso, o caso da função de custo *uniforme* pode ser interpretado como o problema de *maximização do bloqueio de influência*, nesse caso nossos resultados mostram que as métricas de intermediação, percolação e PageRank obtêm desempenhos semelhantes ao escolher o vértice de maior grau da rede. Mostramos que tal semelhança está relacionada à sobreposição dos conjuntos de soluções.

Por outro lado, no caso da função *penalização de grau*, os resultados mostram que as mesmas métricas têm desempenhos opostos, do que no caso *uniforme*. Nossos resultados sugerem que algoritmos cujas soluções se correlacionam com o conjunto dos vértices de maiores graus não apresentam bom desempenho em qualquer um dos cenários propostos. Sobre a função de custo *uniforme*, pode-se concluir que escolher os vértices de maiores graus tem um ganho maior em relação as outras medidas, outro ponto forte para escolha da métrica de grau é o seu tempo de cálculo em comparação com as demais.

Com relação aos trabalhos futuros, recomenda-se elaborar um algoritmo que apresente uma boa solução em relação a solução ótima do problema e que seja escalável, visto que até aqui os algoritmos bons que conhecemos são os de abordagem gulosa. Outro campo que pode ser abordado é verificar o desempenho dos algoritmos gulosos em comparação com as métricas apresentadas nesse trabalho. Ademais, sugere-se também investigar mais sobre as propriedades matemáticas que o problema apresenta, como por exemplo a questão de submodularidade. Por fim temos que novas funções de custo podem ser exploradas, visto que nesse trabalho focamos nas funções de custo de *penalização de grau* e *uniforme*. Outra recomendação é a elaboração de um *survey*, visto que uma das maiores dificuldades encontradas nesse trabalho foi na pesquisa

bibliográfica, uma vez que vários autores trabalham com nomes diferentes ao tratar o problema. Nesse caso um *survey* resultaria em uma boa base para a compreensão dos problemas relatados.

REFERÊNCIAS

[1] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31:211–236, 2017.

[2] N. Arazkhani, M. R. Meybodi, and A. Rezvanian. An efficient algorithm for influence blocking maximization based on community detection. In *5th International Conference on Web Research (ICWR)*, pages 258–263, 2019.

[3] N. Arazkhani, M. R. Meybodi, and A. Rezvanian. Influence blocking maximization in social network using centrality measures. In *5th Conference on Knowledge Based Engineering and Innovation (KBEI)*, pages 492–497, 2019.

[4] Ulrik Brandes. A faster algorithm for betweenness centrality. *The Journal of Mathematical Sociology*, 25, 2004.

[5] Ceren Budak, Divyakant Agrawal, and Amr Abbadi. Limiting the spread of misinformation in social networks. *Proceedings of the 20th International Conference on World Wide Web, WWW 2011*, pages 665–674, 2011.

[6] Wei Chen, Yifei Yuan, and Li Zhang. Scalable influence maximization in social networks under the linear threshold model. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 88–97, 2010.

[7] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucía Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. *ArXiv*, abs/2003.05004, 2020.

[8] Gerard Cornuejols, Marshall L. Fisher, and George L. Nemhauser. Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. *Management Science*, 23(8):789–810, 1977.

[9] Folha de S. Paulo. Agência lupa. `https://piaui.folha.uol.com.br/lupa/`, 2021. Acessado em 26/01/2021.

[10] Universidade de São Paulo. Detector de fake news. `https://nilc-fakenews.herokuapp.com/`, 2018. Acessado em 26/01/2021.

[11] Fernando C. Erd, André L. Vignatti, and Murilo V. G. da Silva. Blocking the spread of misinformation in a network under distinct cost models. *ASONAM'20 - IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2020.

[12] Giorgio Fagiolo. Clustering in complex directed networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 76 2 Pt 2:026107, 2007.

[13] L. A. Wolsey G. L. Nemhauser and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 1978.

[14] Grupo Globo. Fato ou fake. `https://g1.globo.com/fato-ou-fake/`, 2021. Acessado em 26/01/2021.

[15] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, 2008.

[16] Xinran He, Guojie Song, Wei Chen, and Qingye Jiang. Influence Blocking Maximization in Social Networks under the Competitive Linear Threshold Model Technical Report. *arXiv e-prints*, page arXiv:1110.4723, Oct 2011.

[17] David Kempe, Jon Kleinberg, and Eva Tardos. Maximizing the spread of influence through a social network. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 137-146, 2003.

[18] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[19] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Governance in social media: A case study of the Wikipedia promotion process. In *Proc. Int. Conf. on Weblogs and Social Media*, 2010.

[20] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, page 420–429, New York, NY, USA, 2007. Association for Computing Machinery.

[21] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131, 2012. PMID: 26173286.

[22] Michael Ley. The DBLP computer science bibliography: Evolution, research issues, perspectives. In *Proc. Int. Symposium on String Processing and Information Retrieval*, pages 1–10, 2002.

[23] M. Newman. *Networks: An Introduction*. OUP Oxford, 2010.

[24] H. Nguyen and R. Zheng. On budgeted influence maximization in social networks. *IEEE Journal on Selected Areas in Communications*, 31(6):1084–1094, 2013.

[25] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999. Previous number = SIDL-WP-1999-0120.

[26] Canh V. Pham, Hieu V. Duong, Huan X. Hoang, and My T. Thai. Competitive influence maximization within time and budget constraints in online social networks: An algorithmic approach. *Applied Sciences*, 9(11), 2019.

[27] Mahendra Piraveenan, Mikhail Prokopenko, and Liaquat Hossain. Percolation centrality: Quantifying graph-theoretic impact of nodes during percolation in networks. *PLOS ONE*, 8(1):1–14, 2013.

[28] Eduardo Scolese, Fábio Takahashi, and Joelmir Tavares. Disseminação do jornalismo profissional reduz influência de fake news, indica pesquisa. https://www1.folha.uol.com.br/poder/2021/01/disseminacao-do-jornalismo-profissional-reduz-influencia-de-fake-news-indica-pesquisa.shtml, 2021. Acessado em 26/01/2021.

[29] Lovro Šubelj and Marko Bajec. Model of complex networks based on citation dynamics. In *Proceedings of the WWW Workshop on Large Scale Network Analysis*, pages 527–530, 2013.

[30] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[31] Peng Wu and Li Pan. Scalable influence blocking maximization in social networks under competitive independent cascade models. *Computer Networks*, 123, 05 2017.

[32] W. Zhu, W. Yang, S. Xuan, D. Man, W. Wang, and X. Du. Location-aware influence blocking maximization in social networks. *IEEE Access*, 6:61462–61477, 2018.

[33] W. Zhu, W. Yang, S. Xuan, D. Man, W. Wang, X. Du, and M. Guizani. Location-based seeds selection for influence blocking maximization in social networks. *IEEE Access*, 7:27272–27287, 2019.

[34] W. Zhu, W. Yang, S. Xuan, D. Man, W. Wang, and J. Lv. Location-aware targeted influence blocking maximization in social networks. In *2019 28th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–9, 2019.