Concepção de uma Solução Escalável para Maximização de Influência Ciente de Tópicos em Redes Sociais

Daniel Bruno Alves dos Santos

Tese de Doutorado submetida à Coordenação do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Campina Grande - Campus de Campina Grande como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências no Domínio da Engenharia Elétrica.

Área de Concentração: Processamento da Informação

Angelo Perkusich, D.Sc.
Orientador
Hyggo Oliveira de Almeida, D.Sc.
Orientador

Campina Grande, Paraíba, Brasil

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

S237c Santos, Daniel Bruno Alves dos.

Concepção de uma solução escalável para maximização de influência ciente de tópicos em redes sociais / Daniel Bruno Alves dos Santos. – Campina Grande, 2015.

102 f. : il. color.

Tese (Doutorado em Engenharia Elétrica) - Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2015.

"Orientação: Prof. D.Sc. Angelo Perkusich, Prof. D.Sc. Hyggo Oliveira de Almeida".

Referências.

1. Redes Sociais. 2. Maximização de Influência. 3. Tópicos. 4. Escalabilidade. I. Perkusich, Angelo. II. Almeida, Hyggo Oliveira de Almeida. III. Título.

CDU 004.771:316.472.4(043)

"CONCEPÇÃO DE UMA SOLUÇÃO ESCALÁVEL PARA MAXIMIZAÇÃO DE INFLUÊNCIA CIENTE DE TÓPICOS EM REDES SOCIAIS"

DANIEL BRUNO ALVES DOS SANTOS

TESE APROVADA EM 13/11/2015

ANGELO PERKUSICH, D.Sc., UFCG
Orientador(a)

HYGGO OLIVEIRA DE ALMEIDA, D.Sc., UFCG
Orientador(a)

MARIA DE FÁTIMA QUEIROZ VIEIRA, Ph.D., UFCG
Examinador(a)

JAIDILSON JO DA SILVA, D.Sc., UFCG
Examinador(a)

EVANDRO DE BARROS COSTA, D.Sc., UFAL
Examinador(a)

FABRÍCIO BENEVENUTO DE SOUZA, Dr., UFMG
Examinador(a)

CAMPINA GRANDE - PB

À tia Laura.

Agradecimentos

Agradeço primeiramente a Deus, por me proporcionar a vida, a saúde, juntamente com a coragem e a determinação para superar todos os desafios e obstáculos que me foram apresentados.

Aos meus pais e irmã: Jeronimo, Neves e Jeane que me apoiam em todos os momentos. Aos meus cunhados Daniel e Rafael pela força e descontração. Agradeço também aos meus sogros Seu Wolney e Dona Mary, que tanto me ajudaram nesses anos.

À minha esposa, Wania, que sempre tem estado comigo nos momentos felizes e naqueles não tão felizes. Obrigado pela paciência, compreensão e companheirismo de todos esses anos juntos. Se hoje alcancei um sonho foi porque você me apoiou incondicionalmente!

Aos amigos do Residencial Flamingo, pela descontração de todos os momentos.

Também, aos amigos que fiz ainda no mestrado e, principalmente, no doutorado, os quais são provenientes dos mais diversos lugares: Campina Grande, João Pessoa, interior do Ceará, Maceió e Teresina. Em especial aos amigos do Laboratório Embedded, os quais contribuíram direta ou indiretamente com ideias valiosas para o desenvolvimento deste trabalho.

Aos funcionários e funcionárias da COPELE/DEE que sempre foram tão solícitos.

Aos meus orientadores, Angelo Perkusich e Hyggo, pela paciência, orientação recebida e contribuição substancial para a concretização deste trabalho.

Ao CNPq, pelo apoio financeiro.

Resumo

O uso das redes sociais tem demonstrado enorme potencial para a criação, divulgação de informações e formação de opinião. Um dos problemas centrais que tem atraído a atenção de pesquisadores consiste em encontrar um conjunto inicial de usuários que, ao receberem algum incentivo, podem influenciar uma porção substancial da rede social para comprar um produto, adotar uma inovação ou propagar notícias. Este problema é denominado de Maximização de Influência. Embora avanços expressivos tenham sido alcançados desde a definição deste problema, a maior parte dos esforços tem sido concentrada em solucionar limitações de escalabilidade e de como aprender os parâmetros da solução. Como resultado, outros aspectos importantes foram pouco explorados, como, por exemplo, a relação de dependência entre a influência social e os tópicos de interesse dos usuários. Recentemente, essa questão tem sido abordada em um problema denominado de Maximização de Influência baseada em Tópicos, que consiste em encontrar um conjunto inicial de usuários com a habilidade de influenciar uma porção substancial de uma rede social em relação a um tópico específico. Todavia, as soluções propostas não são adequadas para redes sociais de larga escala e precisam incorporar mecanismos para determinar a influência social exercida entre os usuários em relação a cada tópico de interesse. Consequentemente, para estas abordagens, torna-se difícil ou mesmo inviável lidar de forma rápida e eficiente com as mudanças constantes na estrutura das redes sociais. Tal problema é particularmente relevante quando são considerados os tópicos de interesse dos usuários e a influência social que os mesmos exercem uns sobre os outros em cada tópico. Neste trabalho é proposta uma solução escalável baseada em mineração de dados sobre um registro de propagações de informações, com o objetivo de selecionar diretamente o conjunto inicial de usuários influentes em um determinado tópico, sem a necessidade de incorporar uma etapa anterior de aprendizagem de influência social relacionada a esse tópico. Como benefício adicional, o conjunto inicial de usuários obtido possui uma garantia de aproximação em relação à solução ótima. Por fim, é apresentada uma avaliação experimental sobre um conjunto de dados contendo propagações de informações de uma rede social real, onde são obtidas evidências de que a solução proposta mantém um custo-benefício entre escalabilidade e acurácia.

Abstract

The use of social networks has shown great potential for information diffusion and formation of public opinion. One key problem that has attracted researchers' interest is how to find an initial set of users such that, when given an incentive, they might influence a substantial portion of the network to buy a product, adopt an innovation, or spread news. This problem is known as Influence Maximization. Although major improvements have been made since the first solution for this problem was developed, most of these efforts have been concerned on how to solve scalability issues and how to learn the solution parameters. As a result, other key aspects have gained minor interest, such as depending on relationship between social influence and users' topics of interest. Recently, this issue has been addressed as a problem known as Topic-based Influence Maximization, referring to finding a small set of users on a social network that have the ability to influence a substantial portion of users on a given topic. The proposed solutions, however, are not suitable for large-scale social networks and must incorporate mechanisms for determining social influence among users for each topic of interest. Consequently, for these approaches, it becomes difficult or even unfeasible to deal quickly and efficiently with constant changes in the structure of social networks. This problem is particularly relevant when the topics of interest of users and the social influence they exert on each other for every topic are considered together. In this work we propose a scalable solution that makes use of data mining based on an information propagation log, in order to directly select the initial set of influential users on a particular topic without needing to incorporate a previous learning stage of social influence with regard to that topic. As an additional benefit, the targeted seed set also offers an approximation guarantee of the optimal solution. Finally, an experimental evaluation is presented based on datasets containing information propagation data from real social networks where evidence has been found that the proposed solution maintains a trade-off between scalability and accuracy.

Sumário

Lista de Símbolos e Abreviaturas

AIR - *Authoritativeness - Interest - Relevance model*
API - *Application Programming Interface*
BFS - *Breadth-First Search*
CD - *Credit Distribution model*
DAG - *Direct Acyclic Graph*
FTD - *Fecho Transitivo Direto*
FTI - *Fecho Transitivo Inverso*
GCM - *General Cascade Model*
HCD - *Topic-Aware Homophily-based Credit Distribution model*
IC - *Independent Cascade model*
IM - *Influence Maximization*
LDA - *Latent Dirichlet Allocation*
LT - *Linear Threshold model*
LSI - *Latent Semantic Indexing*
MC - *Monte Carlo*
MIS - *Marginal Influence Sort*
RMSE - *Root Mean Squared Error*
SC - *Seed Credits*
TAP - *Topic Affinity Propagation*
TIC - *Topic Independent Cascade model*
TLT - *Topic Linear Threshold model*
UC - *User Credits*
UCC - *User Created Contents*
URL - *Uniform Resource Locator*
WC - *Weighted Cascade model*
VSM - *Vector Space Model*

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Desde a década de 40, os cientistas das áreas de Ciências Sociais têm se interessado pela questão fundamental de como novas tendências, comportamentos e inovações se propagam através das redes sociais. Uma rede social é uma estrutura social constituída por um conjunto de atores (pessoas, grupos, organizações) conectados entre si a partir de relações sociais existentes no mundo real [3]. Alguns exemplos desses relacionamentos que permeiam o dia a dia das pessoas são as relações existentes entre os membros de uma mesma família, as relações de amizade entre pessoas de uma mesma vizinhança, as relações de trabalho entre funcionários de uma empresa, dentre outras.

As primeiras investigações realizadas sobre a difusão e adoção de inovações [4,5] foram focadas nos estudos realizados nas áreas de medicina e agricultura em comunidades rurais nos Estados Unidos [1]. Segundo Rogers [1], uma inovação é uma ideia, prática ou objeto que é percebida como novo, uma nova alternativa, onde *a priori* não se sabe ao certo se é superior à prática original. Por sua vez, difusão de informação é um processo pelo qual uma inovação é comunicada através de canais de comunicação ao longo do tempo entre os membros de um sistema social. Para realizar os estudos sobre difusão de inovações, os pesquisadores precisavam, primeiramente, inferir os relacionamentos sociais existentes entre os membros do sistema social, com base na observação das relações sociais existentes entre eles, a fim de reconstruir as redes sociais das quais eles faziam parte. Conforme apontado por Rogers, essa era obviamente uma tarefa desafiadora e muito custosa de ser realizada, mesmo se fosse considerada uma rede social com centenas ou milhares de usuários. Somente após reconstruírem a rede social, é que os pesquisadores se voltavam para o processo de compreensão do *como* e o *porquê* de algumas inovações (não) serem adotadas em substituição às práticas existentes [1].

Com a popularização do acesso à Internet e o surgimento de diversos serviços sociais na Web, o cenário descrito anteriormente mudou consideravelmente. Nos dias atuais, é possível estudar propagações de informações em redes sociais de larga escala. Hoje, a es-

trutura das redes sociais e as informações de propagação podem ser coletadas diretamente dos mais diversos sistemas de informação, como os serviços de blogs e *microblogging*, serviços de mensagem instantânea, sistemas de recomendação *online* e sistemas de redes sociais *online*. Em particular, os sistemas de redes sociais *online* apresentaram um crescimento acelerado do número de usuários e uma explosão de popularidade sem precedentes. Esses sistemas registram milhões de acessos diários [6, 7] e caracterizam-se por possibilitar aos usuários entrarem em contato com familiares, amigos, colegas de trabalho, bem como, estreitarem as relações sociais existentes no mundo real, a partir da utilização das ferramentas de comunicação disponíveis na própria rede social.

De acordo com Boyd [8], um *sistema de rede social* é um serviço baseado na Web que permite: (1) construir um perfil público ou semi-público dentro de um sistema fechado; (2) construir uma lista de contatos contendo outros usuários do sistema, com os quais um usuário compartilha relacionamentos sociais e; (3) visualizar e navegar nas listas de contatos existentes no perfil de um usuário e no perfil de outros usuários do sistema. Exemplos de sistemas de redes sociais são Facebook[1], LinkedIn[2], Twitter[3], dentre outros. Com a crescente utilização desses sistemas pelos usuários, novas funcionalidades têm se tornado padrão, como por exemplo, serviços de publicação de mensagens e gerenciamento de comentários no próprio perfil do usuário ou no perfil de um amigo; serviços de mensagem instantânea; e ferramentas para armazenamento e compartilhamento de diversos tipos de mídias sociais [9] como vídeos, fotos, músicas, etc.

Ao fornecerem uma diversidade de ferramentas de comunicação aos usuários, os sistemas de redes sociais *online* permitem que os usuários produzam e propaguem informações rapidamente. Por exemplo, no Facebook, uma informação produzida por um usuário pode conter texto, música, foto ou vídeo. Frequentemente, uma combinação desses tipos de mídias pode ser postada no mural dos próprios usuários ou de seus contatos sociais. Uma informação também pode ser proveniente de fontes externas à própria rede social, como portais de notícias, blogs especializados em algum tipo de informação, etc, as quais são inseridas no sistema pelos usuários da rede social. Já a propagação de uma informação é resultante das ações realizadas pelos usuários no sistema. Alguns exemplos são as ações de publicar, comentar, curtir, recomendar algum tipo de informação, dentre outras, que trazem visibilidade àquela informação para os demais usuários.

Com base nessa dinâmica de propagação, de modo semelhante aos canais de mídia de massa, como a televisão, o rádio e os jornais, o uso das redes sociais possibilita aos usuários tornarem-se cientes de eventos que estão ocorrendo em qualquer lugar, seja em nível local, regional, nacional ou mundial. Porém, uma vantagem em favor das redes sociais

[1] `https://www.facebook.com`
[2] `http://www.linkedin.com`
[3] `https://twitter.com`

é que essas informações podem ser personalizadas para cada usuário, isto é, refletindo as preferências e interesses pessoais deles. Outra vantagem é que essas informações são produzidas ou compartilhadas por pessoas que possuem um relacionamento social com o usuário, tornando-as mais relevantes, e elas são recebidas em tempo real. Assim, ao adquirirem ciência de um evento de interesse que está ocorrendo, os usuários podem rapidamente propagar informações relacionadas ao mesmo, opinando sobre o assunto, quer seja com os amigos ou com quaisquer outros usuários da rede social.

Recentemente, os cientistas da área de Análise de Redes Sociais têm se interessado em desenvolver técnicas para analisar o grande volume de dados resultantes das propagações de informações nas redes sociais. Dessa forma, o desenvolvimento de técnicas eficientes para análise de propagações de informações (por exemplo, rastreamento, monitoramento, identificação de tendências, etc) tornou-se uma questão chave para o desenvolvimento dessa área. Como consequência do desenvolvimento dessas técnicas, as propagações de informações têm sido investigadas em diversos contextos, tais como: *marketing* viral [10]; sistemas de recomendação [11–15]; análise de propagações de informações [16]; identificação de especialistas em um domínio de conhecimento [17–20]; identificação de comunidades [21, 22]; predição de *links*, isto é, identificação da estrutura de uma rede social quando esta não é conhecida *a priori*, a partir da observação das cascatas de propagações [23–26]; e propagação de confiança [27], dentre outras.

Em particular, um problema que tem sido amplamente investigado é o de maximizar a propagação de informações entre os usuários de uma rede social. Isto é, fazer com que uma determinada informação seja propagada para a maior quantidade de usuários. A solução deste problema tem forte impacto econômico e apresenta aplicabilidade direta nas áreas de *marketing* [28] e sistemas de recomendação [13, 29].

Dentre as estratégias possíveis para propagar informações para uma grande quantidade de usuários, geralmente a mais rápida e eficiente, segundo Rogers [1], seria realizar uma divulgação em massa para tornar as informações visíveis para todos os usuários do sistema. Na prática, em uma rede social *online*, como por exemplo o Facebook, para que informações estejam disponíveis para todos os usuários, seria necessário exibi-las diretamente nas páginas dos perfis desses usuários.

No entanto, uma divulgação em massa pode apresentar diversos problemas. Primeiro, os usuários que não têm interesse nessas informações irão recebê-las. Segundo, pelo fato dos usuários de uma rede social terem que lidar com milhares de informações recebidas diariamente, eles inconscientemente ignoram ou abstraem a maioria dessas informações, a partir de um processo denominado de atenção seletiva [28]. Terceiro, quando informações são advindas do sistema de rede social, essas informações possivelmente não refletirão os interesses e preferências dos usuários. Além disso, essas informações não conterão

qualquer tipo de avaliação prévia por parte dos amigos do usuário. Por esses motivos, a possibilidade de que as mesmas sejam relevantes para o usuário e, com isso, retenham sua atenção, fazendo com que ele compartilhe-as com seus amigos, é menor do que se as mesmas fossem provenientes da rede de contatos do usuário. Por fim, esta estratégia pode mostrar-se inviável economicamente para um anunciante que esteja interessado em promover um produto na rede social, no caso do custo de divulgação estar diretamente associado à quantidade de usuários existentes.

Na Figura 1.1 é exemplificada a utilização da estratégia de divulgação em massa por parte de uma rede social. Especificamente, é utilizada uma lista dinâmica de produtos divulgados no perfil de um usuário do Facebook, em um espaço de divulgação denominado *Patrocinado*.

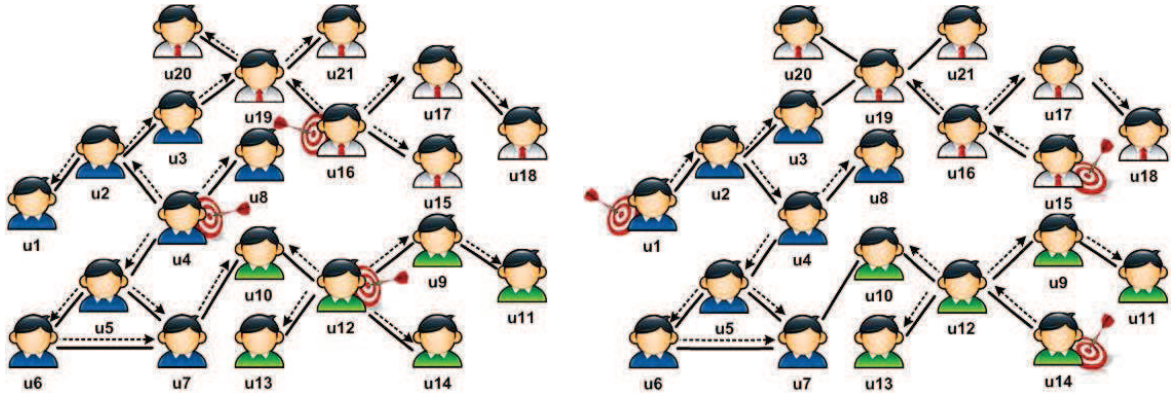
Figura 1.1: Ilustração da utilização da estratégia de divulgação em massa por parte de uma rede social.

Segundo Rogers [1], uma segunda abordagem consiste em explorar as relações sociais dos usuários para propagar informações. Nesta abordagem, primeiro são selecionados alguns usuários da rede social, de acordo com algum critério pré-estabelecido e, em seguida, é solicitado aos mesmos que divulguem a informação para os seus contatos sociais. A grande vantagem desta estratégia em relação à divulgação em massa é que pelo fato das informações serem compartilhadas pelos usuários com seus contatos sociais, essas informações são potencialmente mais relevantes para os usuários envolvidos, minimizando o problema de atenção seletiva [28].

Nesta estratégia, a propagação de informações entre os usuários ocorrerá através de um processo de comunicação denominado *boca a boca*[4]. Em particular, nas redes sociais

[4]Tradução do autor à expressão *word-of-mouth* encontrada na literatura de *marketing* viral.

exemplificadas na Figura 1.2, uma propagação tem início quando os usuários selecionados (re)produzem uma informação e a compartilham com os usuários da sua lista de contatos. Por sua vez, cada um desses usuários pode propagar a mesma informação para os próprios contatos, resultando em uma propagação em cascata. Por exemplo, no Twitter, a propagação de uma notícia relacionada ao lançamento de um novo *smartphone* no mercado é iniciada através da publicação de uma mensagem (*tweet*) da empresa fictícia de tecnologia ACME para alguns usuários selecionados previamente. Por sua vez, esses usuários propagam essa notícia para os seus respectivos seguidores (*followers*) através de republicações (*retweets*) sucessivas da mensagem original. Esse processo de propagação pode ocorrer em vários níveis, onde o primeiro nível é constituído pelos usuários selecionados pela empresa de tecnologia e os demais níveis - dois, três e assim por diante - contêm os seguidores do nível anterior. Esse processo pode perdurar durante um intervalo de tempo indeterminado, resultando em uma longa cascata de propagações da informação original.



(a) Cenário de propagação boca a boca iniciada a partir dos usuários com mais conexões sociais.

(b) Cenário de propagação boca a boca iniciada a partir dos usuários mais influentes.

Figura 1.2: Ilustração de dois cenários de seleção de usuários para propagação boca a boca de informações.

Na rede social ilustrada na Figura 1.2(a), o critério utilizado consiste em selecionar os usuários que têm a maior quantidade de conexões sociais com outros usuários da rede e, portanto, apresentam maior possibilidade de enviarem uma mensagem que consiga alcançar os demais usuários da rede social. O problema dessa estratégia é que a informação pode ser enviada por um usuário que não seja percebido pelos seus contatos como especialista naquele tipo de informação. Por exemplo, na rede social ilustrada na Figura 1.2(a), quando o usuário u_5 compartilha a notícia relacionada ao novo *smartphone* da ACME com os seus amigos, eles podem ignorar a informação compartilhada pelo fato de não reconhecerem o usuário como sendo especialista em dispositivos móveis.

Já na rede social ilustrada na Figura 1.2(b), o critério utilizado consiste em selecionar os usuários com base na influência social que eles exercem sobre a rede social. Essa estratégia tem sido explorada no contexto de *marketing* viral e tem como premissa a ideia

de que se os usuários mais influentes forem selecionados para promover um produto, eles poderão ativar uma longa cascata de propagações com base em sua influência.

Segundo Rashotte [30], a influência social é definida como a mudança de pensamentos, sentimentos, atitudes ou comportamentos de um indivíduo, que são resultantes da interação social com outro indivíduo ou grupos de indivíduos. Com base nessa definição, supõem-se que, quando os usuários observam seus amigos realizarem uma ação e, após um intervalo de tempo, eles também decidem executá-la, esses usuários estão sendo influenciados pelos seus amigos. Um exemplo prático seria o de um usuário que decide assistir a um vídeo no Youtube[5] após ler recomendações dos seus amigos nas redes sociais.

A principal diferença entre as estratégias de seleção discutidas é que embora os usuários mais influentes não apresentem necessariamente a maior quantidade de conexões sociais, eles podem ser percebidos pelos seus contatos como especialistas em determinados domínios de conhecimento. Com isso, quando esses usuários são selecionados para divulgar informações de seu domínio, eles possuem maior possibilidade de persuadirem os seus contatos sociais a propagarem essas informações na rede social.

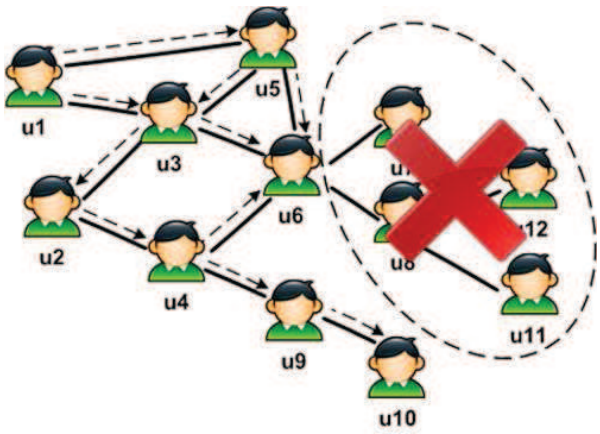


Figura 1.3: Ilustração da interrupção da propagação de informações em uma rede social. Caso o usuário u_6 deixe de compartilhar a informação recebida, todos os usuários acessíveis a partir dele deixarão de receber essa informação.

Portanto, ao selecionar um usuário é necessário considerar a relevância das informações que estão sendo propagadas e o nível de influência que o usuário exerce em sua rede social. Esses requisitos não funcionais são importantes, pois, conforme ilustrado na Figura 1.3, caso as informações recebidas não sejam relevantes ou os usuários selecionados não consigam influenciar os seus contatos, provavelmente não ocorrerá o compartilhamento dessas informações com esses contatos, ocasionando uma interrupção do processo de propagação. Como consequência, potenciais interessados nessas informações, que podem estar

[5] `http://www.youtube.com`

acessíveis a partir do usuário que falhou na tentativa de propagar essas informações para os seus contatos, deixarão de recebê-las.

As estratégias de propagação descritas anteriormente evidenciam a importância da etapa de seleção dos usuários. A seleção dos usuários mais influentes em sistemas de redes sociais de larga escala é um desafio que tem recebido grande atenção nos últimos anos. Essa etapa tem-se mostrado fundamental no processo de propagação de informações e, consequentemente, para obtenção de longas cascatas de propagação.

É no contexto de selecionar os usuários mais influentes em uma rede social *online* de larga escala, capazes de maximizar a propagação de informações para os usuários que se interessem pelas mesmas, minimizando o recebimento de informações irrelevantes, que se insere o presente trabalho.

1.1 Problemática

Para ajudar a compreender o problema de selecionar os usuários mais influentes que maximizam a propagação de uma informação em uma rede social *online*, será apresentado um cenário de aplicação, no contexto de *marketing* viral no qual o mesmo ocorre. No cenário exemplificado na Figura 1.4, um anunciante de uma empresa de tecnologia tem por objetivo divulgar um novo *smartphone* em um sistema de rede social *online*, de modo que o produto anunciado seja divulgado para a maior quantidade possível de usuários interessados em sua aquisição.

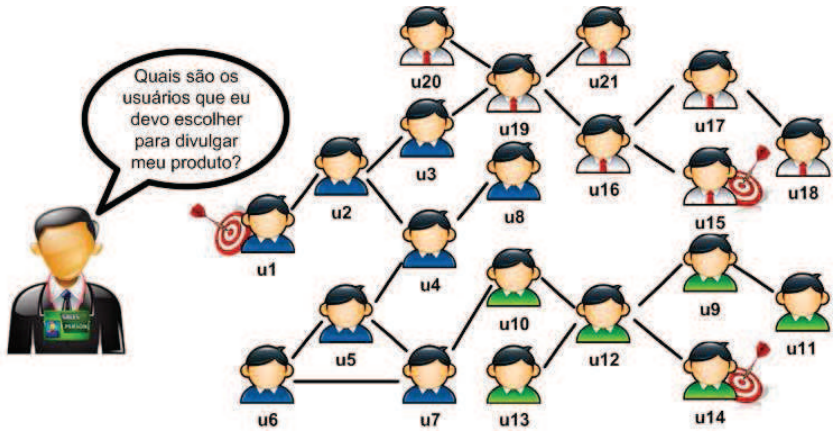


Figura 1.4: Ilustração do problema de negócio no contexto de *marketing* viral.

Considere que o provedor do sistema de rede social cobre um determinado valor sobre cada anúncio divulgado diretamente pelo anunciante. Considere ainda que o anunciante possui restrições no seu orçamento, não podendo pagar mais do que uma quantidade limitada de usuários para propagarem seu anúncio. Logo, se quiser atingir o seu objetivo, o

anunciante deverá selecionar os "melhores usuários" da rede social e convencê-los a divulgar o produto entre os seus amigos, amigos dos amigos, amigos dos amigos dos amigos, e assim por diante. Nesse caso, a divulgação do produto dependerá exclusivamente da força das relações sociais [31] existentes entre os usuários, ou seja, do nível de influência de quem está propagando a informação e também do nível de interesse em relação ao produto anunciado.

Com base na descrição anterior, quais são os usuários da rede social que o anunciante deverá selecionar para divulgação do seu produto, respeitando as restrições existentes, a fim de que o anúncio seja divulgado para a maior quantidade de usuários que tenham interesse no produto?

Domingos e Richardson [32, 33] foram os primeiros a abordarem o problema de selecionar o conjunto inicial de usuários no contexto de *marketing* viral e apresentar uma solução probabilística para o mesmo. Nesse sentido, os autores propuseram algoritmos para avaliar não apenas o valor de um usuário na rede social (i.e. o interesse do usuário na aquisição de um produto), mas também o valor da rede de amigos do usuário (i.e. a quantidade de amigos que poderiam ser influenciados por aquele usuário para aquisição de um produto).

Posteriormente, Kempe *et al.* [34], focando nos modelos de difusão *Independent Cascade (IC)* e *Linear Threshold (LT)*, estudaram esse problema como um problema de otimização discreta.

No modelo IC, o processo de difusão é iniciado com um conjunto inicial de nós (i.e. usuários) ativos A_0, e o processo continua em passos discretos de acordo com a seguinte regra aleatória: Quando um nó v é ativado no passo t, a ele é dada uma única chance de ativar cada um dos seus vizinhos u que esteja inativo. Ele terá sucesso com uma probabilidade $p_{v,u}$. Se v obtiver sucesso, então u se tornará ativo no passo $t+1$. Obtendo sucesso ou não, v não poderá tentar ativar u novamente em qualquer rodada subsequente. Esse processo continuará até que nenhuma nova ativação seja possível [34].

No modelo LT, um nó u é influenciado por cada vizinho v de acordo com um peso $p_{v,u}$, tal que o somatório dos pesos dos vizinhos que possuem uma aresta incidente em u é menor ou igual a 1. A dinâmica deste processo ocorre como segue: Cada nó u escolhe um limiar θ_u uniformemente e de modo aleatório do intervalo $[0, 1]$. Este limiar corresponde à fração dos amigos de u que devem se tornar ativos para que u também seja ativado. Em seguida, dado um conjunto inicial de nós ativos, denotado por A_0 (com todos os outros nós inativos), o processo de difusão acontece de modo determinístico em passos discretos: no passo t todos os nós que foram ativados no passo $t-1$ permanecem ativos, e todos os nós u, cujo somatório dos pesos dos vizinhos ativos é pelo menos θ_u, são ativados. Novamente, esse processo continuará até que nenhuma nova ativação seja possível [34].

Kempe *et al.* [34] também formalizaram esse problema em seu trabalho seminal como problema de Maximização de Influência (IM): dado um modelo de propagação m (por exemplo, IC ou LT) e um conjunto inicial $S \subseteq V$, o número esperado de nós ativos no fim do processo é denotado por $\sigma_m(S)$. O problema de Maximização de Influência refere-se a encontrar o conjunto $S \subseteq V$, $|S| = k$, tal que, $\sigma_m(S)$ é máximo.

Na formalização de Kempe *et al.*, o problema de Maximização de Influência requer como parâmetros de entrada dois tipos de dados. O primeiro é um grafo direcionado $G = (V, E)$, onde V é o conjunto de usuários e E é o conjunto das relações ou ligações sociais existentes entre os usuários, respectivamente. O segundo parâmetro de entrada são as probabilidades sobre todas as arestas do grafo, onde essas probabilidades indicam o nível de influência existente entre quaisquer dois usuários no grafo G.

Kempe *et al.* também provaram que, utilizando os modelos IC e LT, o problema IM é NP-Difícil. Entretanto, os autores também demonstraram que quando uma função $\sigma_m(S)$ é monótona e submodular, então há um algoritmo *Greedy* que em cada iteração acrescenta ao conjunto de nós S, o nó que provê o maior ganho marginal, e produz uma solução com garantia de aproximação de $(1 - 1/e - \epsilon)$, para qualquer $\epsilon > 0$, em relação à solução ótima [34]. Em particular, uma função é monótona quando $\sigma_m(S) \leq \sigma_m(T)$, toda vez que $S \subseteq T$ (isto é, quando um elemento é adicionado a um conjunto, ele não causa uma diminuição em $\sigma_m(S)$). Ainda, uma função é submodular quando $\sigma_m(S \cup \{w\}) - \sigma_m(S) \geq \sigma_m(T \cup \{w\}) - \sigma_m(T)$, toda vez que $S \subseteq T$ (i.e. o ganho marginal obtido ao adicionar um elemento ao conjunto S é pelo menos igual ao ganho marginal obtido ao adicionar o mesmo elemento ao superconjunto de S).

Uma limitação chave dessa solução para o problema IM é a ineficiência do algoritmo *Greedy*, quando utilizado sobre redes sociais de larga escala. Isso se deve ao fato de, em cada iteração do algoritmo, serem necessárias execuções de simulações Monte Carlo (MC) para selecionar o nó que provê o maior ganho marginal. Tipicamente, em cada execução de simulações MC, são realizadas 10 mil amostragens sobre um grafo G (cada amostra também é um grafo). Em seguida, é escolhido um nó $u \in V - S$, onde V é o conjunto de vértices e S é o conjunto inicial. Então, sobre cada amostra, é computado o tamanho da propagação produzida por $S \cup \{u\}$. O tamanho da propagação é dado pelo número de nós presentes na amostra que são alcançáveis a partir do conjunto inicial (incluindo os elementos do próprio conjunto). O valor médio do tamanho da propagação de todas as amostras é utilizado para estimar a propagação esperada produzida pelo conjunto inicial.

De modo geral, são necessárias $O(nk)$ execuções de Simulações MC para selecionar todos os k-elementos do conjunto inicial de usuários, onde n é o número de usuários em uma rede social. Assim, dependendo do tamanho da rede social e do parâmetro k especificado no problema, podem ser necessários vários dias para obter uma solução

para o problema IM [35]. Desde então, com o objetivo de melhorar o desempenho do algoritmo *Greedy* e torná-lo computacionalmente mais eficiente, vários pesquisadores têm contribuído com o desenvolvimento de otimizações [36–44] e heurísticas [38,39,45], [46–51].

Outra limitação existente na solução proposta por Kempe *et al.* [34] é a indisponibilidade das probabilidades existentes sobre as arestas do grafo. Enquanto as informações sobre a estrutura das redes sociais *online* estão amplamente disponíveis e podem ser facilmente coletadas, como por exemplo, através de processos de *web crawling* [52], as probabilidades existentes sobre as arestas do grafo não são conhecidas *a priori*. Por esse motivo, a maioria dos trabalhos que abordam o problema de Maximização de Influência, que utilizam os modelos de propagação *Independent Cascade*, *Weighted Cascade* e *Linear Threshold*, fazem a suposição de que as probabilidades das arestas são fornecidas como entrada do problema, ou ainda, simplificam o processo de aprendizagem dessas probabilidades ao utilizarem uma das estratégias de atribuição enumeradas a seguir:

- **Modelo Weighted Cascade**. Para cada aresta $(v, u) \in E$, é atribuído um valor $1/D_{in}(v)$, onde $D_{in}(v)$ representa o grau de entrada de um vértice v (i.e. o número de vizinhos que possuem uma aresta orientada em direção a v);

- **Modelo de Trivalência**. Para cada aresta $(v, u) \in E$, é atribuída uma probabilidade uniforme, de modo aleatório, com base em um dos valores do conjunto de constantes $\{0{,}1; 0{,}01; 0{,}001\}$;

- **Modelo Uniforme**. Para cada aresta $(v, u) \in E$, é atribuída uma probabilidade com valor constante $p = 0{,}01$.

Com base em uma análise comparativa entre as abordagens de atribuição de influência, no trabalho de Goyal *et al.* [35] são fornecidas evidências de que os métodos enumerados anteriormente impactam negativamente na qualidade dos conjuntos iniciais de usuários encontrados pelas soluções baseadas no algoritmo *Greedy*. Em particular, os conjuntos iniciais encontrados por esses métodos possuem pouca ou nenhuma interseção entre si. Por outro lado, os autores também demonstraram que métodos que incorporam uma etapa de aprendizagem da influência social a partir de dados reais, coletados dos registros de propagações ocorridas nas redes sociais *online*, produzem resultados com melhor qualidade. Esse resultado reforça a necessidade de serem desenvolvidas técnicas de aprendizagem de influência social utilizando o histórico de propagações. Nesta linha de pesquisa, destacam-se os trabalhos de Saito, Nakano e Kimura [53]; Saito *et al.* [54] e; Goyal, Bonchi e Lakshmanan [55].

Recentemente, Goyal *et al.* [35] propuseram uma abordagem alternativa para encontrar o conjunto de k-usuários que maximizam a influência em uma rede social, a partir

de dados de propagações reais. Nessa abordagem, os autores introduziram um novo parâmetro de entrada ao problema original: os *registros de propagações de ações*. Nesses registros são armazenadas informações sobre quem executou uma determinada ação em um dado intervalo de tempo. Além disso, os autores propuseram um modelo de Distribuição de Créditos, que é alimentado por um processo de mineração direta sobre os registros de propagações de ações. Esse modelo é utilizado para agregar créditos diretos e transitivos aos usuários, dados pelos seus contatos diretos e indiretos, pelo fato desses usuários terem influenciado os contatos na execução de cada ação. Ao utilizar dados de propagações reais, não é necessário incorporar uma etapa de aprendizagem das probabilidades das arestas e, tampouco, executar simulações Monte Carlo. Dessa forma, é possível realizar a construção de uma solução baseada no algoritmo *Greedy*, que seja capaz de selecionar rapidamente um conjunto inicial com k-usuários. Adicionalmente, resultados de uma avaliação experimental evidenciam que a solução é escalável para redes sociais de larga escala. Todavia, uma limitação dessa abordagem é que a mesma tem como premissa a disponibilidade de registros de propagações sobre diversas ações para todos os usuários, uma premissa que não é válida na prática [56].

Apesar dos avanços significativos das abordagens anteriores para o problema de Maximização de Influência, todas elas têm como base a seguinte suposição: os indivíduos possuem a habilidade de influenciar seus amigos com a mesma intensidade, em qualquer assunto ou tópico. Entretanto, tal suposição não modela com exatidão a realidade, uma vez que contradiz algumas teorias sociológicas sobre o comportamento social coletivo dos indivíduos, como por exemplo a teoria de Granovetter [57]. De acordo com a teoria de Granovetter, a influência social entre indivíduos distintos depende dos múltiplos tópicos de interesse desses indivíduos. Além disso, os indivíduos exercem influência social distinta sobre os seus amigos em cada tópico. Por exemplo, a influência social que um indivíduo u exerce sobre outro indivíduo v, no tópico futebol, provavelmente será diferente quando o tópico referir-se à política. Ainda, considerando um tópico como política, a influência que o indivíduo u exerce sobre o indivíduo v provavelmente será diferente daquela que v exerce sobre u.

Reconhecendo a necessidade de considerar a dependência entre a influência social e os tópicos de interesse dos usuários, vários pesquisadores [22], [58–60] começaram a estudar essas questões, a partir da observação dos comportamentos dos indivíduos nos sistemas de redes sociais.

Liu *et al.* [58] e Tang *et al.* [59] foram os primeiros a formalizarem em seus trabalhos um Modelo de Tópicos aplicado a grafos para redes sociais de larga escala, denominado de *Topical Affinity Propagation* (TAP). A partir da utilização de TAP, os autores mostraram que é possível modelar a influência social entre cada par de usuários por tópicos. Além

disso, para cada tópico é possível especificar um valor distinto para a influência social entre os usuários. Embora o grafo de influência gerado a partir de TAP satisfaça as especificidades enumeradas anteriormente, tal solução foi utilizada apenas em problemas para encontrar especialistas ou construir subgrafos de influência por tópicos. Portanto, os autores não abordam o problema de Maximização de Influência.

Por sua vez, Zang *et al.* [60] foram os primeiros a abordarem o problema de Maximização de Influência baseado em Tópicos. Em particular, os autores consideram as preferências dos usuários sobre diferentes tópicos. Essas preferências podem ser calculadas a partir da escolha entre duas técnicas alternativas: *Latent Semantic Indexing* (LSI) e *Vector Space Model* (VSM). Para ambas as técnicas é obtido como resultado um vetor de preferências por tópico, onde são armazenados valores reais que representam a importância de cada tópico para o usuário. Esse vetor é passado como parâmetro de entrada para uma versão adaptada do algoritmo *Greedy*, sendo utilizado para ponderar a influência por tópico de interesse de cada usuário.

Barbieri *et al.* [22] também estudaram o problema de Maximização de Influência baseado em Tópicos. Em seu trabalho, os autores estenderam os modelos de propagação IC e LT, tornando-os cientes de tópicos, os quais foram denominados de *Topic-aware Independent Cascade* (TIC) e *Topic-Aware Linear Threshold* (TLT), respectivamente. Os autores também introduziram um novo modelo de propagação de influência, o qual eles denominaram de AIR (*Authoritativeness* - *Interest* - *Relevance*), pois além de considerar a influência social, em tal modelo também são considerados outros fatores, tais como a autoridade de um usuário sobre o tópico, o interesse de um usuário sobre o tópico e a relevância de um item para um tópico. Apesar desses modelos utilizarem o conceito de registro de propagações para aprender os parâmetros dos modelos, ainda são utilizadas Simulações Monte Carlo para selecionar o conjunto inicial de usuários. Por fim, diferentemente dos modelos TIC e TLT, o conjunto inicial de usuários obtido com base no modelo AIR não oferece nenhuma garantia de aproximação em relação à solução ótima.

Após o surgimento das primeiras soluções para o problema de Maximização de Influência baseado em Tópicos, alguns autores [61–64] observaram que essas soluções não eram adequadas para encontrar rapidamente em uma rede social, os usuários mais influentes para um determinado tópico de interesse em tempo real.

Barbieri *et al.* [61], propuseram um método baseado em similaridade, denominado de INFLEX, para encontrar de forma rápida os usuários que maximizam a propagação de informações relacionadas a um dado tópico. Para isso, os autores utilizaram o algoritmo *Greedy* para pré-processar (fase *offline*) alguns conjuntos iniciais para determinados tópicos e, em seguida, construíram um índice desses conjuntos iniciais para tópicos específicos. A ideia chave é que ao ser realizada uma consulta contendo tópicos já indexados (fase *on-*

line), o INFLEX possa determinar rapidamente dentre os tópicos já indexados, qual deles é o mais similar em relação à consulta realizada e o conjunto inicial referente a esse tópico seja utilizado para responder a consulta.

Inspirados no INFLEX, Chen *et al.* [62] introduziram o algoritmo *Marginal Influence Sort* (MIS), que pode ser utilizado para pré-processar os valores de influência marginal dos usuários para cada tópico. Assim, o objetivo dos autores é tornar ainda mais rápido o cálculo dos valores de influência marginal e aumentar a eficiência do algoritmo INFLEX no processo de seleção dos usuários presentes no conjunto inicial.

Quase que simultaneamente, Chen *et al.* [63] e Li *et al.* [64] também propuseram soluções que utilizam técnicas de amostragem do conjunto inicial para vários tópicos, para acelerar o processamento de seleção dos usuários presentes no conjunto inicial. Em comum, os autores desses trabalhos afirmam que suas soluções são escaláveis para redes sociais reais e apresentam garantias de aproximação em relação à solução ótima.

Uma grande limitação enfrentada pelas soluções anteriores, é que há uma enorme quantidade de combinações possíveis de consultas a serem armazenadas. Desse modo, tais soluções necessitarão de uma grande quantidade de espaço em disco para armazenar uma quantidade exponencial de combinações possíveis de tópicos. Além disso, para cada tópico será necessária a realização de uma etapa de treinamento dos modelos, a fim de aprender os valores das probabilidades para cada par de usuários. Por fim, outras soluções [61, 62] não apresentam garantias de aproximação em relação à propagação obtida pela solução ótima.

As redes sociais são ambientes muito dinâmicos, nos quais a estrutura do grafo social está em constante alteração, seja devido à entrada e/ou saída de usuários, ou simplesmente, devido ao surgimento e/ou término de relacionamentos sociais entre os usuários. Adicionalmente, tais sistemas também são ambientes altamente favoráveis ao compartilhamento de informações e troca de opiniões, que podem resultar em mudanças nos interesses e no comportamento dos usuários.

Quando considerada em conjunto com as características de dinamicidade descritas anteriormente, a escala dos sistemas de redes sociais, medida pela quantidade total de usuários e de conexões sociais registradas no sistema, contribui fortemente para o aumento da complexidade de encontrar de forma eficiente um conjunto inicial de usuários que seja uma solução para o problema de Maximização de Influência baseado em Tópicos. Dessa forma, considerando tais desafios, é necessário que soluções para esse problema possam lidar de forma eficiente com essas características. Em particular, tais soluções devem refletir rapidamente as mudanças na estrutura de uma rede social, nos tópicos de interesse dos usuários e também permitir que a influência social, exercida pelos usuários em cada tópico, seja (re)aprendida rapidamente, enquanto mantém a qualidade do conjunto inicial

de usuários encontrado.

De modo geral, as soluções para o problema de Maximização de Influência baseado em Tópicos incorporam uma etapa de aprendizagem da influência social relacionada a cada tópico, armazenadas sobre as arestas do grafo, e precisam executar simulações Monte Carlo para encontrar o conjunto inicial. Entretanto, conforme discutido previamente, essas soluções não são escaláveis e, portanto, não contemplam as características descritas anteriormente. Por sua vez, as soluções que (re)aprendem as informações sobre a influência social a partir do histórico de propagações e que não necessitam de simulações Monte Carlo para selecionar os usuários que farão parte do conjunto inicial, apresentam-se como soluções mais promissoras, pois são capazes de lidar com a escala, a dinamicidade inerente aos sistemas de redes sociais e ainda conseguem encontrar conjuntos iniciais eficientes. Todavia, as soluções existentes [35] não consideram os tópicos de interesse dos usuários.

Analisando as soluções existentes para o problema de Maximização de Influência baseado em Tópicos, não foram identificadas soluções que lidem com as características de dinamicidade e escala de uma rede social e que ainda encontre um conjunto inicial de usuários com garantia de aproximação em relação à solução ótima. Este é o problema abordado neste trabalho.

1.2 Objetivo Geral

Neste trabalho, tem-se como objetivo principal a concepção de uma solução escalável para maximizar a propagação de informações em redes sociais *online*, com base na influência social e nos tópicos de interesse dos usuários. Mais especificamente, a solução proposta, a partir de um grafo representando uma rede social, um histórico de propagações de informações, um tópico de interesse e um parâmetro k representando a quantidade de usuários a serem encontrados, permite:

- Inferir dinamicamente os tópicos de interesse dos usuários a partir de dados de propagações reais;
- Inferir o nível de influência social entre os usuários, considerando tópicos de interesses similares;
- Minerar diretamente um conjunto de k-usuários que maximizam a propagação de informações na rede social por tópico de interesse dos usuários.

1.3 Objetivos Específicos

Considerando o objetivo geral deste trabalho, pode-se dividi-lo nos seguintes objetivos específicos:

1. Definir um modelo para representar o histórico de propagações realizadas pelos usuários em uma rede social;

2. Definir um modelo para aprendizagem de influência social por tópicos de interesse dos usuários, a partir de registros de propagações reais;

3. Especificar e desenvolver uma solução para encontrar os k-usuários que maximizam a propagação de informações em redes sociais, de acordo com a influência social exercida pelos usuários em cada tópico de interesse e as características dos conteúdos propagados na rede social;

4. Realizar uma validação técnica da solução. Especificamente, pretende-se avaliar a acurácia da predição do modelo, a qualidade do conjunto inicial encontrado, o tamanho das propagações obtidas a partir do conjunto inicial e o tempo necessário para encontrar o conjunto inicial, em relação às soluções encontradas no estado da arte.

1.4 Relevância

As redes sociais *online* proveem grandes oportunidades para estudar o problema de Maximização de Influência baseado em Tópicos, pois elas conectam uma enorme quantidade de pessoas e armazenam enormes volumes de informações relacionadas às estruturas sociais e a dinâmica de comunicação dentro do sistema. Entretanto, elas também apresentam desafios para resolução desse problema. Além de possuírem uma larga escala, as redes sociais *online* representam estruturas sociais complexas e muito dinâmicas, o que significa que a solução para esse problema necessita ser muito eficiente e escalável.

Desde quando Domingos e Richardson [32,33] iniciaram o estudo sobre como encontrar os usuários mais influentes em redes sociais e, principalmente, após Kempe *et al.* [34] formalizarem o problema de Maximização de Influência como um problema de otimização discreta, o problema de Maximização de Influência baseado em Tópicos tem se mantido um tópico relevante nas áreas de Mineração de Dados e Análise de Influência Social. Esse aspecto é corroborado pela imensa quantidade de artigos encontrados em veículos relevantes da área.

No que diz respeito à relevância do trabalho, em se tratando de um problema de tese, os seguintes aspectos foram levados em consideração, os quais reforçam o caráter de relevância desta tese.

O primeiro aspecto é o desafio técnico. O problema de encontrar um conjunto inicial contendo *k-usuários* para o problema de Maximização de Influência baseado em Tópicos é um problema NP-Difícil. Desse modo, a especificação de uma solução eficiente, que considere a dinamicidade e a escala dos sistemas de redes sociais existentes, e que ao mesmo tempo encontre um conjunto inicial que ofereça uma garantia de aproximação em relação à solução ótima, é um problema técnico desafiador.

O segundo aspecto é a consistência teórica. A solução proposta foi construída utilizando como formalismos a Teoria dos Grafos e a Teoria dos Conjuntos. Especificamente, os conceitos utilizados pelo modelo de Distribuição de Créditos baseado em Tópicos foram fundamentados nas suposições encontradas na Teoria sociológica sobre o comportamento social coletivo de Granovetter [57] e também nos conceitos do arcabouço teórico de Homofilia [65].

O terceiro aspecto é a contribuição científica. Diversos trabalhos relacionados foram estudados antes da concepção da solução apresentada nesta tese. A partir deste estudo, identificou-se o problema enunciado anteriormente e, em seguida, foram elencadas as possíveis soluções para o problema, o que culminou com a definição deste trabalho. Até o momento da escrita deste documento, não foram encontrados trabalhos com as características aqui propostas, o que reforça o caráter de originalidade e a contribuição científica.

Por fim, o problema de Maximização de Influência baseado em Tópicos também tem sido estudado em outros contextos, tais como: detecção de epidemias [66]; detecção de eventos e de novas tendências populares [67]; detecção de boatos, opiniões ou sentimentos negativos, acerca de um tópico/produto/marca, que estão propagando-se rapidamente em uma rede social [68,69].

1.5 Organização do Documento

O restante deste documento está organizado da seguinte forma:

- No Capítulo 2, apresenta-se a base conceitual para o entendimento do restante do documento. Em particular, devido à natureza multidisciplinar do problema abordado, o arcabouço teórico apresentado inclui conceitos relacionados às áreas de Ciência da Computação e Sociologia. Esses conceitos incluem: Teoria dos Grafos, Redes Sociais, Influência Social, Homofilia, Difusão de Inovações e Modelos de Propagação.

- No Capítulo 3, apresenta-se uma solução escalável para o problema de Maximização

de Influência baseado em Tópicos.

- No Capítulo 4, descreve-se a realização de um projeto experimental, cujo intuito é demonstrar a validade técnica da solução proposta no Capítulo 3.
- No Capítulo 5, são apresentadas as conclusões e as perspectivas futuras decorrentes desta tese.

Capítulo 2

Fundamentação Teórica

Neste capítulo são apresentadas as terminologias e os conceitos fundamentais para o entendimento do restante deste trabalho.

2.1 Redes Sociais

Nas próximas seções são apresentadas as diferentes visões que estão relacionadas ao conceito de Redes Sociais.

2.1.1 Perspectiva das Ciências Sociais

O estudo das redes sociais é multidisciplinar e, portanto, permeia diversas áreas do conhecimento tais como as Ciências Sociais (Antropologia, Sociologia, Comunicação, etc) e Tecnologia da Informação, dentre outras.

Para as Ciências Sociais, as redes sociais são utilizadas como forma de representar e de compreender os relacionamentos sociais existentes na sociedade. Dessa forma, esse conceito é utilizado com o objetivo de estudar como os indivíduos estão organizados na sociedade e como eles se relacionam, a partir da observação da interação individuo-individuo ou individuo-grupos. Como exemplo, esse conceito pode ser utilizado para reconstruir as relações de parentesco, ou ainda, as relações existentes dentro de uma empresa, entre organizações diferentes, etc.

2.1.2 Perspectiva da Análise de Redes Sociais

A área de Análise de Rede [3, 70] está preocupada em estudar as redes sociais do ponto de vista das propriedades de sua estrutura. Essa abordagem é fundamentalmente matemática e utiliza a Teoria dos Grafos [71] para representar os elementos das redes sociais.

Uma rede social é definida como um conjunto de dois elementos: atores (pessoas, instituições ou grupos) e suas conexões (interações ou laços sociais) [3]. Esses elementos são representados como grafos.

Dentro desta perspectiva, Degenne e Forsé [70] explicam que a "Análise de Rede é um conjunto de métodos recente para o estudo sistemático das estruturas sociais".

Na área de Análise de Redes Sociais, a ideia de rede é utilizada enquanto uma ferramenta de análise dos relacionamentos entre as pessoas, seus elos pessoais e entre as organizações, no contexto que elas se inserem. Desse modo, são analisadas propriedades relacionadas à estrutura da rede social, tais como a densidade, a clusterização e a presença de grupos. Há também um foco sobre os atores, a partir de um conjunto de propriedades relacionadas aos aspectos estruturais da rede social, na qual esses atores estão inseridos. Dentre essas propriedades, considera-se, por exemplo, a quantidade de conexões que o usuário possui (grau de conexão); se o usuário está localizado em uma posição central na rede (grau de centralidade); se o mesmo faz parte de vários grupos e; se ele exerce o papel de intermediador (grau de intermediação), na troca de informações entre os grupos dos quais faz parte.

2.1.3 Perspectiva Tecnológica

O conceito de redes sociais também está relacionado a uma perspectiva tecnológica, enquanto sistema de informação na Web. Em um trabalho recente, Kim *et al.* [9] elaboraram uma taxonomia para classificar as redes sociais em dois tipos de sistemas: *redes sociais online* e *mídias sociais*. Esses dois sistemas são antes de tudo *web sites sociais*. A seguir será apresentada a definição de *web sites* sociais e, em seguida, as definições de redes sociais *online* e mídias sociais.

Os *Web sites sociais* são definidos como aqueles sites que possibilitam às pessoas formarem *comunidades online* e compartilharem conteúdos produzidos pelos usuários (*User created contents* - UCC) [9]. Nessa definição, as *pessoas* são quaisquer usuários que utilizam a Internet, ou ainda, usuários de uma organização (por exemplo, Universidade, empresa, governo, etc). Já uma *comunidade* pode ser formada, por exemplo, pelas redes de amigos, de conhecidos, de colegas de trabalho, dentre outras. Essas redes fazem parte do que se denomina "mundo real" (também utiliza-se o termo *offline*) e passam a fazer parte da rede observada no que se denomina de mundo *online*. Alguns exemplos de UCC são as informações que constituem o perfil dos usuários, bem como, fotos, imagens, vídeos, atualizações das atividades dos usuários, texto (por exemplo, postagens e comentários), dentre outros. O *compartilhamento* de um UCC inclui, no mínimo, a postagem, a visualização e o comentário do UCC por um usuário (incluindo o próprio usuário que criou o UCC).

Por sua vez, as redes sociais *online* são definidas como sites da Web que permitem às pessoas estarem conectadas com outras pessoas em comunidades *online* [9]. No entanto, uma definição mais precisa de uma rede social *online* é aquela utilizada por Boyd *et al.* [8]. De acordo com Boyd [8], uma *rede social online* é um serviço baseado na Web que permite: (1) construir um perfil público ou semi-público dentro de um sistema fechado; (2) construir uma lista de contatos contendo outros usuários do sistema, com os quais um usuário compartilha relacionamentos sociais e; (3) visualizar e navegar nas listas de contatos existentes no perfil de um usuário e no perfil de outros usuários do sistema. Alguns exemplos de sistemas de redes sociais são o Facebook[1], LinkedIn[2], Twitter[3], dentre outros.

Por fim, as *mídias sociais* são definidas como sites da Web que permitem às pessoas compartilharem UCCs [9]. Alguns exemplos de mídias sociais são o Fickr[4], Youtube[5], Instagram[6], dentre outros.

Com a crescente utilização desses web sites sociais por parte dos usuários, novas funcionalidades têm se tornado padrão, como por exemplo, serviços de publicação de mensagens e gerenciamento de comentários no próprio perfil do usuário ou no perfil de um amigo; serviços de mensagem instantânea e; ferramentas para armazenamento e compartilhamento de diversos tipos de mídias sociais, tais como vídeos, fotos, imagens, músicas, etc. Desse modo, conforme apontado por Kim *et al.* [9], a diferenciação outrora existente em relação aos termos redes sociais *online* e mídias sociais passa a não mais existir.

2.2 Difusão de Informações

Segundo Rogers [1], difusão é o processo pelo qual uma inovação é comunicada por meio de certos canais no decorrer do tempo entre os membros de um sistema social. É um tipo especial de comunicação que está preocupada com a propagação de mensagens que são percebidas como ideias novas. Já a comunicação é um processo no qual os participantes de um sistema social criam e compartilham informações uns com os outros, com o objetivo de alcançar um entendimento mútuo. De modo geral, o processo de difusão é caracterizado por um certo grau de incerteza e risco que o indivíduo sente a respeito da inovação. Conforme aponta Rogers [1], um indivíduo pode reduzir tais preocupações quando ele obtém informação sobre a inovação.

Os principais elementos no processo de difusão de novas ideias são: a *inovação* que é

[1] `https://www.facebook.com`
[2] `http://www.linkedin.com`
[3] `https://twitter.com`
[4] `https://www.flickr.com`
[5] `http://www.youtube.com`
[6] `http://instagram.com`

comunicada através de *canais de comunicação* ao longo do *tempo* entre os membros de um *sistema social*. Esses elementos serão explicados nas próximas seções.

2.2.1 Inovação

Uma *inovação* é uma ideia, prática, ou objeto percebido como novo por um indivíduo ou outra unidade de adoção [1]. Uma inovação não precisa necessariamente ser algo inédito. Ela pode surgir simplesmente como uma alternativa a uma prática já existente. Conforme apontado por Rogers [1], *a priori* quem irá adotar a inovação não sabe se a mesma é melhor ou pior em relação às práticas já consolidadas. Nesse caso, a falta de informação é um dos fatores que pode levar os indivíduos a resistirem em relação à adoção do "novo".

2.2.2 Canais de Comunicação

Um canal de comunicação é um meio pelo qual mensagens são enviadas de um indivíduo para outro [1]. Existem dois tipos básicos de canais de comunicação: as *mídias de massa* (por exemplo, televisão, jornais, rádio, Internet, etc) e os *canais interpessoais*. As mídias de massa são mais efetivas para criar o conhecimento coletivo sobre as inovações, uma vez que a informação é enviada para todos os indivíduos de um sistema social simultaneamente [1]. Já os canais interpessoais são mais efetivos em formar e mudar atitudes em relação a uma nova ideia. Desse modo, eles são canais mais eficazes para influenciar a decisão de adotar ou rejeitar uma nova ideia [1].

Conforme apontado por Rogers, a maioria dos indivíduos avalia uma inovação não apenas com base no conhecimento científico de especialistas, mas principalmente por meio de avaliações subjetivas de pessoas que apresentam uma relação de proximidade com o usuário (por exemplo, pessoas da família, amigos, colegas de trabalho) e de outros indivíduos que já adotaram a inovação (denominados de usuários). É a avaliação positiva ou negativa desses usuários a respeito da inovação, que faz com que a incerteza sobre o novo diminua e que influencia no processo de decisão sobre adotar ou rejeitar a inovação.

2.2.3 Tempo

Segundo Rogers [1], o tempo está envolvido no processo de difusão nos seguintes momentos: (1) durante o processo de difusão de inovações, (2) no momento em que o indivíduo adere à inovação e (3) na taxa de adoção da inovação.

O processo de decisão de inovação compreende as etapas nas quais um indivíduo tem o primeiro contato com a inovação (*conhecimento*), até o instante em que ele forma uma atitude em relação a mesma (*persuasão*), passando pela sua decisão em adotar ou rejeitar a inovação (*decisão*). Quando um indivíduo decide adotar uma inovação, ele implementa a

mesma (*implementação*) e, após observar os resultados obtidos, ele confirma sua adotação de modo definitivo (*confirmação*). Segundo Rogers, um indivíduo procura informações em vários níveis no processo de tomada de decisão, com o objetivo de diminuir os riscos e a incerteza a respeito das consequências advindas da adoção da inovação.

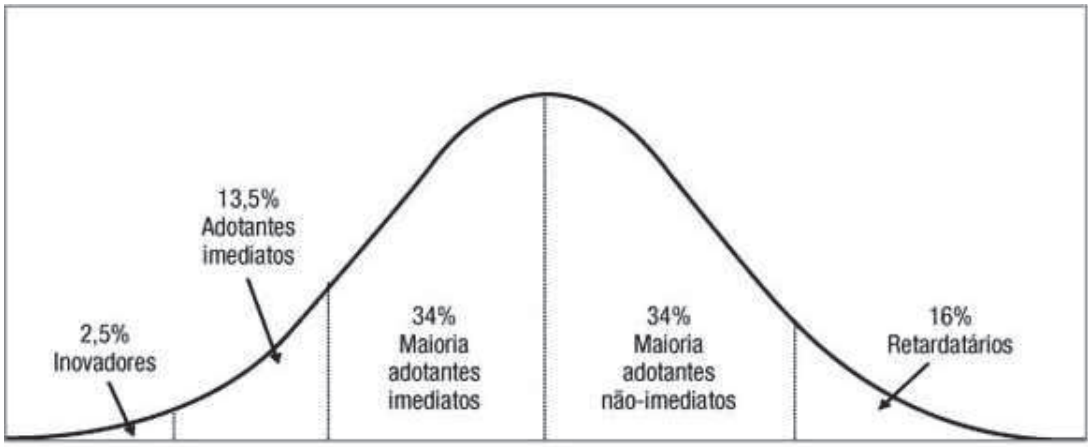


Figura 2.1: Formato da Curva de Adoção de inovações. Fonte: adaptado de [1]

.

O tempo de inovação está relacionado ao momento no qual um indivíduo tende a adotar uma inovação. Tipicamente, o processo de adoção é caracterizado graficamente conforme ilustrado na Figura 2.1. Analisando as informações presentes na Figura 2.1, percebe-se claramente a presença de cinco regiões bem destacadas. Cada uma dessas regiões está associada com a velocidade em que o usuário adota uma inovação. Desse modo, com base na velocidade em que um indivíduo adota uma inovação em relação aos outros membros do sistema social, ele pode ser classificado em uma das cinco categorias de adoção descritas a seguir:

- ***Inovadores (Innovators)***: são indivíduos que estão sempre a procura de novas ideias e, tipicamente, são os primeiros em um sistema social a adotarem uma inovação. Esses indivíduos conseguem lidar com as incertezas relacionadas aos benefícios trazidos pela inovação;

- ***Adotantes imediatos (Early adopters)***: são indivíduos que assumem um papel de liderança em sua rede social. Tipicamente, são bem informados em relação a alguma área de conhecimento (por exemplo, dispositivos móveis), sendo por tanto, respeitados pelos outros indivíduos do sistema social e vistos como especialistas;

- ***Maioria adotantes imediatos (Early majority)***: são indivíduos que tendem a adotarem novas ideias antes da média dos membros de um sistema social. E, que proveem conectividade com os outros membros do sistema social;

- ***Maioria adotantes não-imediatos (Late majority)***: são indivíduos que adotam as inovações após a média dos membros de um sistema social o fazerem. Normalmente, são indivíduos céticos em relação aos benefícios trazidos por uma inovação;
- ***Retardatários (Laggards)***: são os últimos indivíduos a adotarem uma inovação.

Por fim, a taxa de adesão está relacionada com a velocidade na qual uma inovação é adotada pelos membros de um sistema social.

2.2.4 Sistema Social

Um sistema social é um conjunto de unidades inter-relacionadas que estão engajadas no problema de atingir um objetivo comum. Um sistema tem uma estrutura, definida com base nos padrões de conexões que as unidades apresentam. A estrutura de um sistema social pode facilitar ou impedir que a difusão de inovação ocorra. Por exemplo, se os indivíduos estiverem isolados em pequenos grupos, sem nenhuma conexão que tenha acesso direto aos demais grupos do sistema, uma inovação não se difundirá para esses outros grupos. Por outro lado, se a partir das conexões existentes no sistema, for possível acessar qualquer indivíduo, então, a possibilidade que uma inovação se espalhe para uma grande quantidade de indivíduos será maior.

2.2.5 Difusão de Informações e Redes Sociais

Em decorrência das interações que ocorrem entre os atores de um sistema social, são formados os laços sociais que vão conectar os atores nas redes sociais. O conteúdo dessas interações pode ser utilizado no processo de caracterização de um determinado laço social. Granovetter [31] classifica os laços sociais como *fortes* e *fracos*. Os laços fortes seriam aqueles caracterizados pelo grande investimento de tempo, pela criação de intimidade, de confiança e de reciprocidade entre os atores. Os laços fracos, ao contrario, possuem menor quantidade desses elementos, caracterizando, relações menos profundas, não traduzindo proximidade ou intimidade entre os indivíduos. São caracterizadas por apresentarem apenas relações esparsas, com muitas trocas sociais [72]. Um laço forte, por exemplo, seria aquele que se tem com um amigo. Um laço fraco, por outro lado, seria aquele que caracteriza o relacionamento com um conhecido.

No caso das difusões de informações em redes sociais, as mesmas são observadas através das conexões. Granovetter [31] apontou que os laços fracos teriam extrema importância nesse padrão, pois seriam esses laços que manteriam a rede interconectada e que seriam responsáveis pelo fluxo de informações atingir pontos diferentes da rede.

Ao analisar alguns padrões de interação existentes nas redes sociais, Recuero [72] identificou que diferentes tipos de redes sociais poderiam ser formadas de acordo com o tipo de interação existente entre os indivíduos. De acordo com Recuero [72], essas interações são classificadas em *mútuas* ou *reativas*.

Quando os relacionamentos sociais são formados através de interação mútua (por exemplo, a interação constante entre os indivíduos), eles tendem a originar redes sociais com menos indivíduos, entretanto, mais densas e que são caracterizadas por reciprocidade, suporte social e confiança mútua. Essas redes estão associadas ao conceito de laços fortes de Granovetter [31].

Por sua vez, as interações reativas estariam relacionadas a associação dos indivíduos a comunidades ou grupos com os quais eles se identificam. Por exemplo, um indivíduo no Facebook pode se associar a uma comunidade denominada de "Mochileiros", pelo fato de ter interesse nesse tipo de atividade. Uma vez associado a comunidades ou grupos, os indivíduos tendem a não interagir frequentemente com os outros membros, pois eles estão mais interessados em obter informações de interesse. Segundo Recuero [72], esse tipo de interação forma redes sociais com uma grande quantidade de indivíduos, com uma grande quantidade de conectores (indivíduos que servem de ligação ou ponte com outros indivíduos da rede social). Essas redes estão associadas com a ideia de laços fracos de Granovetter. Como ressalta Granovetter [31], esses laços são eficientes no transporte de informação, mas não tão eficientes na construção de suporte social e confiança. Por fim, para Recuero [72], redes sociais com essas estruturas tendem a difundir informações voltadas para a reputação. E as informações compartilhadas têm valor em sua novidade, ou seja, na primazia da divulgação junto a uma determinada rede interativa.

2.3 Influência Social e Homofilia

Outros dois conceitos fundamentais que estão relacionados à difusão de informações são: *influência social* e *homofilia*.

Segundo Rashotte [30], a *influência social* é definida como a mudança de pensamentos, sentimentos, atitudes ou comportamentos de um indivíduo, que são resultantes da interação social com outro indivíduo ou grupos de indivíduos. Esse conceito pode ainda ser definido como a força que um indivíduo A (isto é, influenciador) exerce sobre outro indivíduo B que introduz uma mudança de comportamento e/ou opinião de B.

Por sua vez, Mcpherson *et al.* [65] definem a *homofilia* (também conhecido como seleção) como a tendência de um indivíduo formar relacionamentos sociais com pessoas similares e, portanto, realizar as mesmas ações. Por exemplo, as pessoas tendem a formar laços de amizade com outras pessoas que: tenham os mesmos interesses, sejam do mesmo

gênero, tenham a mesma idade, sejam da mesma família, trabalham ou estudam no mesmo local, moram na mesma cidade, sejam da mesma classe social.

Vários pesquisadores [73–77] têm encontrado evidências da existência de uma correlação entre os conceitos de influência social e de homofilia na formação dos relacionamentos sociais entre os indivíduos nas redes sociais. Enquanto alguns pesquisadores [74, 76] focaram em distinguir os efeitos da influência social e da homofilia, outros [75, 77] têm investigado os efeitos da retroalimentação que ambos os fatores apresentam um sobre o outro. De acordo com Anagnostopoulos *et al.* [74], fatores externos também estão correlacionados aos dois conceitos anteriores. Por exemplo, o ambiente em que os indivíduos estão localizados pode ser considerado um fator externo. A justificativa é que dois indivíduos podem ter se tornado amigos ou realizado uma mesma ação, devido a, coincidentemente, morarem na mesma cidade e de estarem tirando fotos de um mesmo local simultaneamente. E, mais tarde, essas fotos podem ser publicadas em uma mídia social como, por exemplo, o Instagram.

Os conceitos de influência social e de homofilia são importantes, pois eles podem ser utilizados para modelar o comportamento dos indivíduos nas redes sociais. Assim, com base em tal modelo pode-se tentar predizer o comportamento futuro desses indivíduos. Por esse motivo, esses conceitos têm sido frequentemente utilizados em muitas aplicações de mineração de dados, tais como sistemas de recomendação (utiliza o conceito de similaridade) e *marketing* viral (utiliza o conceito de influência social) e principalmente, no problema de Maximização de Influência.

2.3.1 Medição de Influência Social

Apesar da importância da noção de influência social, ainda não há um consenso na literatura sobre qual é a melhor forma de medir a influência de um determinado usuário. Desse modo, alguns autores [78, 79] têm experimentado métricas relacionadas à popularidade do usuário (número de seguidores ou amigos de um usuário), popularidade do conteúdo (número de *retweets* ou citações), ou mesmo, uma combinação das métricas anteriores. Outros autores [18, 80] têm optado por utilizar métricas de influência que exploram variações do algoritmo *PageRank* [81], pois entendem que métricas relacionadas a popularidade do usuário na rede social (por exemplo, número de seguidores no Twitter, número de amigos no Facebook) não necessariamente implicam em influência social [79].

Kwak *et al.* [78] compararam três medidas diferentes de influência: o número de seguidores (*followers*), valor de *PageRank*, e o número de *retweets*. Em seguida, Cha *et al.* [79] também compararam três diferentes tipos de métricas de influência: número de seguidores, número de *retweets* e número de menções. Por fim, Weng *et al.* [18], propuseram *TwitterRank* para quantificar a influência de usuários no Twitter. Em particular,

TwitterRank utiliza uma métrica baseada no valor do algoritmo *PageRank*, que leva em consideração o tópico de interesse do usuário. Em comum, os autores dos três trabalhos anteriores, observaram que o *ranking* (posição) de um usuário dentro da lista dos usuários mais influentes é dependente da métrica utilizada. Uma vez que, os usuários que obtiveram alta pontuação em uma das métricas avaliadas, não necessariamente obtiveram alta pontuação nas demais métricas.

Inspirados no *TwitterRank*, Silva *et al.* [80] propuseram *ProfileRank*, um modelo de difusão de informação que permite a medição da relevância do conteúdo e da influência do usuário baseado em *Random walks* sobre um grafo bipartido representando usuários e conteúdos. O princípio básico explorado pelo método *ProfileRank* é que conteúdo relevante é criado e propagado por usuários influentes e que usuários influentes propagam conteúdos relevantes. Esse princípio é semelhante àquele utilizado na formulação de algoritmos como *PageRank* e *HITS* [82].

Em outra linha de pesquisa, Bakshy *et al.* [83] propôs um método para quantificar a influência de um usuário (denominado de inicializador), a partir da observação do número de usuários sucessivos que republicaram uma URL postada por esse usuário. Especificamente, os autores propuseram um modelo baseado em árvore de regressão que permite predizer a influência de um indivíduo com base nas seguintes características: (1) atributos do inicializador (número de seguidores, número de amigos, número de *tweets*, data em que os usuários tornaram-se amigos) e; (2) histórico das atividades realizadas pelo inicializador (média do número de republicações de uma URL por parte dos amigos diretos de um usuário e a média do número total de usuários que republicaram a URL).

Por fim, algumas empresas na Web, tais como *Klout*[7], *Peer Index*[8] e *Influencer50*[9], tem desenvolvido e disponibilizado seus próprios índices para medir a influência social *online* dos usuários em sistemas de redes sociais tais como *Twitter*, *Facebook* e *LinkedIn*.

2.3.2 Modelos de Difusão de Influência

Nesta seção são descritos os modelos de difusão de influência comumente encontrados na literatura de Análise de Influência Social [84].

Heurísticas

O modelo mais simples para mensurar a influência de cada vértice é utilizando uma heurística. A seguir são descritas algumas das heurísticas comumente utilizadas:

[7] `https://klout.com/home`
[8] `https://www.brandwatch.com/peerindex-and-brandwatch/`
[9] `http://influencer50.com/`

1. *High-degree*. O conjunto inicial é escolhido de acordo com o grau de cada vértice ($deg(v)$). A ideia chave é que os vértices com a maior quantidade de vizinhos exerceriam uma maior influência sobre seus vizinhos diretos. Na literatura de sociologia, essa estratégia também é conhecida como centralidade de grau;

2. *Low-distance*. Outra heurística comumente utilizada para medir a influência, é a posição do vértice em relação ao centro da rede. A ideia chave é escolher como elementos do conjunto inicial, os vértices que estejam mais próximos (caminhos mais curtos) dos demais vértices da rede. Desse modo, nessa estratégia utiliza-se a intuição de que os indivíduos têm uma maior possibilidade de serem influenciados por aqueles que estejam mais proximamente relacionados a eles [85];

3. *Degree Discount*. Também conhecida como *SingleDiscount*, esta heurística foi introduzida no trabalho de Chen *et al.* [38]. A ideia básica dessa abordagem é que se o vértice u foi selecionado como elemento do conjunto inicial, então, quando v for avaliado como candidato a novo elemento do conjunto inicial, com base no seu grau ($deg(v)$), a aresta (v,u) não deverá ser contabilizada. Mais especificamente, para um vértice v com $deg(v)$ vizinhos, dos quais t_v já foram selecionados para o conjunto inicial, então o desconto no grau de v deve ser $2t_v + (deg(v) - t_v)t_v p$, onde p (por exemplo, $p = 0{,}01$) é a probabilidade de propagação.

Modelos Clássicos

Nos modelos de difusão descritos nesta seção, cada usuário possui um status associado: ativo ou inativo. O status dos usuários que devem ser selecionados para o conjunto inicial S (por exemplo, os usuários que devem ser escolhidos em uma campanha de *marketing*) é visto como ativo. Por sua vez, o status dos demais usuários é visto como inativo. O problema de Maximização de Influência é estudado com o uso dessa dinâmica baseada no status. Inicialmente, todos os usuários são considerados inativos. Então, os usuários escolhidos são ativados e eles podem influenciar seus amigos (vértices vizinhos) e também torná-los ativos.

Linear Threshold Model

No modelo *Linear Threshold* (LT), um nó u é influenciado por cada vizinho v de acordo com um peso $p_{v,u}$, tal que o somatório dos pesos dos vizinhos que possuem uma aresta incidente em u é menor ou igual a 1. A dinâmica deste processo ocorre como segue.

Cada nó u escolhe um limiar θ_u uniformemente e de modo aleatório do intervalo $[0, 1]$. Este limiar corresponde à fração dos amigos de u que devem se tornar ativos para que u também seja ativado. Em seguida, dado um conjunto inicial de nós ativos,

denotado por A_0 (com todos os outros nós inativos), o processo de difusão acontece de modo determinístico em passos discretos: no passo t todos os nós que foram ativados no passo $t-1$ permanecem ativos, e todos os nós u, cujo somatório dos pesos dos vizinhos ativos é pelo menos θ_u, estarão ativados no passo $t+1$. Esse processo continuará até que nenhuma nova ativação seja possível [34].

Independent Cascade Model

No modelo *Independent Cascade (IC)*, o processo de difusão é iniciado com um conjunto inicial de nós ativos A_0, e o processo continua em passos discretos de acordo com a seguinte regra aleatória. Quando um nó v é ativado no passo t, a ele é dada uma única chance de tentar ativar cada um dos seus vizinhos u que estejam inativos. Ele terá sucesso com uma probabilidade $p_{v,u}$. Se v obtiver sucesso, então u se tornará ativo no passo $t+1$. Obtendo sucesso ou não, v não poderá tentar ativar u novamente em qualquer rodada subsequente. Esse processo continuará até que nenhuma nova ativação seja possível [34].

Weighted Cascade Model

O modelo *Weighted Cascade* é um caso especial do modelo *Independent Cascade*. Especificamente, para cada aresta $(u,v) \in E$, a probabilidade de que o vértice u ative seu vizinho v, denotada por $p_{u,v}$, é $p_{u,v} = 1/deg(v)$.

General Cascade Model

Em um trabalho posterior, Kempe *et al.* [86] introduziram um modelo de propagação que generaliza os modelos LT e IC, denominado de *General Cascade Model* (GCM), bem como, especificaram um conjunto de regras de transformação entre esses modelos.

2.4 Teoria dos Grafos

A Teoria dos Grafos [71] é uma parte da matemática aplicada que se dedica a estudar as propriedades dos diferentes tipos de grafos.

Nesta seção são apresentados alguns conceitos relacionados à Teoria dos Grafos. O leitor já familiarizado com tal conteúdo pode omitir a sua leitura sem prejuízo à compreensão do restante deste documento.

2.4.1 Definições Básicas

Um grafo é uma abstração que permite especificar relacionamentos entre pares de objetos presentes em uma coleção [87]. Esses objetos podem ser pessoas, cidades, empresas, entre

outros. Um relacionamento entre pares de objetos pode representar amizade, parentesco, relação de trabalho entre pessoas, conexões entre cidades, associações entre uma pessoa e uma empresa, etc.

Formalmente, um grafo G é definido por $G = (V, E)$, onde V representa um conjunto finito não-vazio de vértices ou nós e; E, representa um conjunto finito de arestas (v, u), onde $v, u \in V$ [88].

Laço

Um laço (*loop*) em um grafo $G = (V, E)$ é uma aresta $e = (v, u) \in E$, onde $v = u$. Em outras palavras, em um grafo G há um laço, sempre que em uma das arestas o vértice de origem e o de destino forem os mesmos [89, 90].

Ordem

A ordem de um grafo G, denotada por $n(G)$, é dada pela cardinalidade do conjunto de vértices, ou seja, $n(G) = |V|$ [89].

De forma análoga, o número de arestas em um grafo G, denotado por $m(G)$, é dada por $m(G) = |E|$. A quantidade mínima de arestas em um grafo G é $m = 0$, e a quantidade máxima de arestas pode ser calculada como: $\binom{n}{2} = \frac{n(n-1)}{2} \leq n^2$.

Adjacência e Incidência

Em um grafo $G = (V, E)$ dois vértices v e u são adjacentes (vizinhos), se existe uma aresta $e = (v, u) \in E$ em G. Neste caso, a aresta $e = (v, u) \in E$ é dita incidente aos vértices v e u [87].

Com base na definição anterior, a vizinhança de um vértice v é composta por todos os vértices u com os quais v está conectado diretamente por meio de uma aresta.

Ainda, para grafos direcionados, o conceito de vizinhança pode se ser especializado em dois novos conceitos:

- *Antecessor:* um vértice u é antecessor do vértice v, se há um arco que parte de u e chega em v;
- *Sucessor:* um vértice u é sucessor do vértice v, se há um arco que parte de v e chega em u.

Grau

O grau de um vértice v (também denominado de valência), denotado por $deg(v)$, é calculado pelo número de arestas que são incidentes ao vértice v, dado por $|E(v)|$. Sendo zero,

o grau mínimo de um vértice qualquer e $n(G) - 1$, o seu grau máximo. Um vértice de grau zero é um vértice isolado no grafo [87].

Como exemplo, considere um grafo não-direcionado $G = (V,E)$, onde $V = \{1, 2, 3, 4\}$ e $E = \{(1, 2),(1, 3),(2, 3),(3, 4)\}$. Logo, para os vértices 1 e 3, o grau desses vértices será $deg(1) = 2$ e $deg(3) = 3$.

Ainda, para grafos direcionados, o conceito de grau pode ser especializado em dois novos conceitos:

- *Grau de entrada:* o grau de entrada de um vértice v, denotado por $deg_{in}(v)$, é representado pelo número de arestas que são incidentes ao vértice v, onde o vértice v faz parte do destino. Isto é, $e = (u,v) \in E$;
- *Grau de saída:* o grau de saída de um vértice v, denotado por $deg_{out}(v)$, é representado pelo número de arestas incidentes ao vértice v, onde o vértice v faz parte da origem. Isto é, $e = (v,u) \in E$.

Como exemplo, considere um grafo direcionado $G = (V,E)$, onde $V = \{1, 2, 3, 4\}$ e $E = \{(1, 2), (1, 3), (2, 3), (3, 4)\}$. Logo, para os vértices 1 e 3, o grau de entrada desses vértices será $deg_{in}(1) = 0$ e $deg_{in}(3) = 2$. De forma análoga, o grau de saída desses vértices será $deg_{out}(1) = 2$ e $deg_{out}(3) = 1$.

2.4.2 Formas de Representação

Nesta seção são descritas as formas de representação de um grafo.

Representação Gráfica

Graficamente, os vértices de um grafo são representados através de pontos ou círculos. Já as arestas são representadas através de linhas conectando pares de vértices [87].

Matriz de Adjacência

Um grafo pode ser representado através de uma matriz de adjacência. Nessa representação, os vértices são associados às linhas e às colunas de uma matriz [87, 91].

Seja $A = [a_{ij}]$ uma matriz $n \times n$, onde n é o número de vértices existentes no grafo $G = (V, E)$. Uma matriz de adjacência A pode ser construída da seguinte forma:

$$a_{ij} = \begin{cases} 1 & \text{se } \exists\, (i, j) \in E, \\ 0 & \text{caso contrário.} \end{cases}$$

Como exemplo, considere a matriz de adjacência apresentada a seguir:

	1	2	3	4
1	0	1	1	0
2	1	0	1	0
3	1	1	0	1
4	0	0	1	0

O grafo $G = (V, E)$, que corresponde à representação da matriz de adjacência apresentada anteriormente, é ilustrado na Figura 2.2.

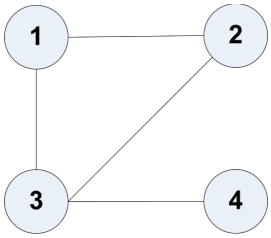


Figura 2.2: Ilustração de um grafo construído a partir de uma matriz de adjacência.

Matriz de Incidência

Uma forma alternativa de representar um grafo é através de uma matriz de incidência. Nessa representação, os vértices são associados às linhas e as arestas são associadas às colunas de uma matriz [87, 91].

Seja $B = [b_{ij}]$ uma matriz $n \times m$, onde $n = |V|$ e $m = |E|$, representando um grafo orientado $G = (V, E)$. Uma matriz de incidência B pode ser construída da seguinte forma:

$$b_{ij} = \begin{cases} 1 & \text{se vértice } i \text{ incide sobre a aresta } j, \\ 0 & \text{caso contrário.} \end{cases}$$

Como exemplo, considere a matriz de incidência apresentada a seguir:

	e_1	e_2	e_3	e_4
1	0	1	1	0
2	1	0	1	0
3	1	1	0	1
4	0	0	1	0

O grafo $G = (V, E)$, que corresponde à representação matricial apresentada anteriormente, é ilustrado na Figura 2.3.

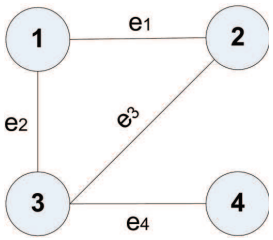


Figura 2.3: Ilustração de um grafo construído a partir de uma matriz de incidência.

Lista de Adjacência

Uma forma eficiente, em termos de consumo de memória, de representar um grafo é por meio de uma lista de adjacência. A ideia básica é associar a cada vértice uma lista de vértices adjacentes [87,91].

Nessa representação, os vértices pertencentes ao conjunto de vértices V de $G = (V, E)$ são associados a um vetor de tamanho n, onde $n = |V|$. Adicionalmente, cada vértice pode estar associado a uma lista de vértices adjacentes [87,91].

Na Figura 2.4, é ilustrado um grafo G representado por meio de uma lista de adjacência.

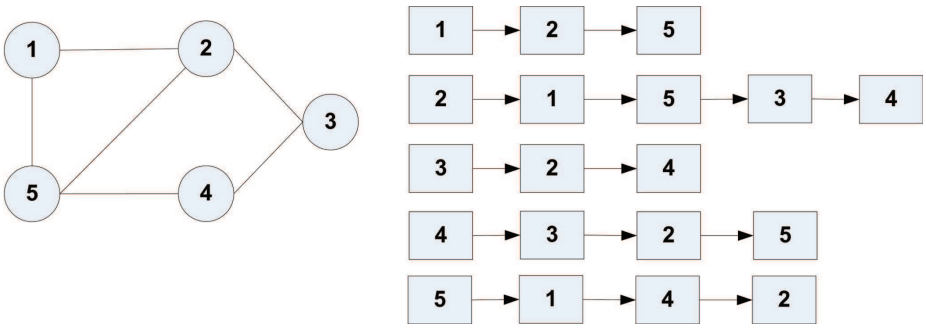


Figura 2.4: Ilustração de um grafo construído a partir de uma lista de adjacência.

2.4.3 Tipos de Grafos

Nesta seção são apresentados alguns tipos de grafos definidos na Teoria dos Grafos.

Grafo Não-Direcionado

Um grafo é dito *não direcionado* quando não existe um sentido (orientação) em suas arestas. Dessa forma, cada aresta é um par não-ordenado $(v, u) = (u, v)$. Isto significa que as relações representadas nas arestas são simétricas [89].

Seja, por exemplo, o grafo $G = (V, E)$ dado por:

- $V = \{\text{p} \mid \text{p é uma pessoa}\}$;
- $E = \{(v, u) \mid v \text{ é amigo de } u\}$.

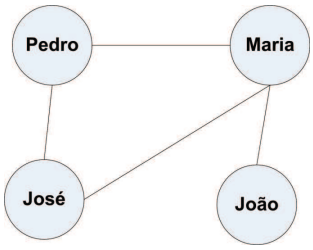


Figura 2.5: Ilustração de um grafo não-direcionado.

Considere, como exemplo, o grafo ilustrado na Figura 2.5, que representa a relação *v é amigo de u*. Esta relação é simétrica, pois se *v é amigo de u*, então *u é amigo de v*.

Grafo Direcionado

Um grafo é dito *direcionado* ou *dígrafo* quando existe um sentido (orientação) em suas arestas. Nesse caso, as conexões entre os vértices são chamadas de arcos. Um sentido é representado graficamente através de uma seta, onde a origem representa o vértice de origem e a extremidade contendo a ponta da seta representa o vértice de destino [89].

Considere, o grafo definido por:

- $V = \{\text{p} \mid \text{p é uma pessoa da família Silva}\}$
- $E = \{(\text{v,u}) \mid v \text{ é pai/mãe de } u\}$

A relação de parentesco, ilustrada na Figura 2.6, não é simétrica. Pois, se *v é pai/mãe de u*, não é o caso que *u é pai/mãe de v*. Há, portanto, uma orientação na relação, com um correspondente efeito na representação gráfica de G.

Quando um grafo orientado não contém ciclos simples em suas arestas (acíclico), ele é denominado de *Direct Acyclic Graph* (DAG) [92].

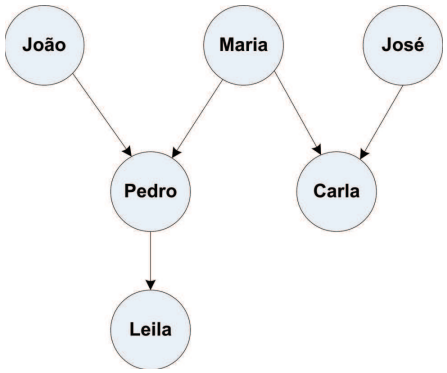


Figura 2.6: Ilustração de um grafo direcionado.

Grafo Ponderado

Um grafo $G = (V, E, w)$ é dito *ponderado* quando cada aresta tem um peso associado a ela, representado na forma $w(v, u)$. O peso é dado por uma função $w : E \to \mathbb{R}$ [87].

Na Figura 2.7, é ilustrado um grafo não direcionado representando cidades conectadas por meio de estradas. Cada par de cidades está localizada a uma distância em quilômetros (Km) uma da outra. O relacionamento, entre os elementos descritos anteriormente, pode ser modelado como um grafo não direcionado ponderado, representado por $G = (V, E, w)$, onde $V = \{\text{A, B, C, D, G}\}$ e $E = \{\text{(A, B, 1015), (A, C, 716), (B, C, 586), (B, D, 429),(B, G, 72), (C, D, 434)}\}$.

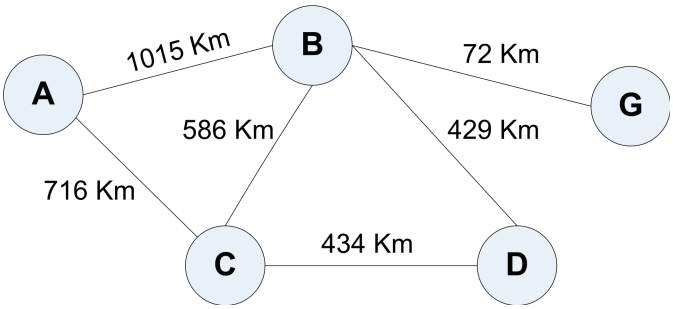


Figura 2.7: Ilustração de um grafo ponderado.

Grafo Simples

Um grafo é denominado *simples* quando: (i) suas arestas não apresentam orientação, (ii) nenhum de seus vértices está relacionado consigo mesmo (isto é, não possui laços) e (iii) não existem arestas paralelas entre os vértices [87].

Na Figura 2.8 é ilustrado um exemplo de grafo simples.

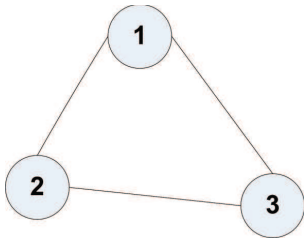


Figura 2.8: Ilustração de um grafo simples.

Grafo Regular

Um grafo é denominado *regular* quando todos os vértices contém o mesmo grau (denotado por r). Nesse caso, estes grafos são designados por *r-regular*. Em um grafo regular, a quantidade de arestas é dada por: $|E| = \frac{nr}{2}$ [87].

O grafo simples ilustrado na Figura 2.8 é um exemplo de grafo 2-regular.

Grafo Completo

Um grafo é denominado *completo* quando há uma aresta conectando cada par de vértices no grafo. Isto é, quando todos os vértices tem grau máximo. Estes grafos são designados por k_n, onde $n = |V|$. Em um grafo completo, a quantidade de arestas é dada por: $|E| = \frac{n(n-1)}{2}$, correspondendo a todas as possíveis escolhas de pares de vértices [87].

Na Figura 2.9, é ilustrado um exemplo de grafo completo.

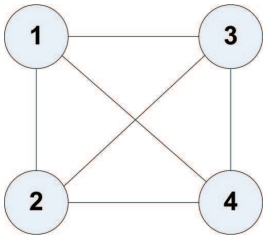


Figura 2.9: Ilustração de um grafo completo k_4.

Grafo Bipartido

Um grafo $G = (V, E)$ é dito ser *bipartido* quando seu conjunto de vértices V puder ser particionado em dois subconjuntos V_1 e V_2, tal que toda aresta de G une um vértice de V_1 a outro de V_2 [89].

Como exemplo, sejam os conjuntos H = {h | h é homem} e M = {m | m é mulher}.

Onde:

- $V = H \cup M$;

- $E = \{(v, w) \mid (v \in H \text{ e } w \in M) \text{ ou } (v \in M \text{ e } w \in H)\}$. A semântica da relação representada pelas arestas (v, w) indica que v *estudou com* w.

Na Figura 2.10, é ilustrado o grafo bipartido $G = (V,E)$ que representa o relacionamento descrito anteriormente.

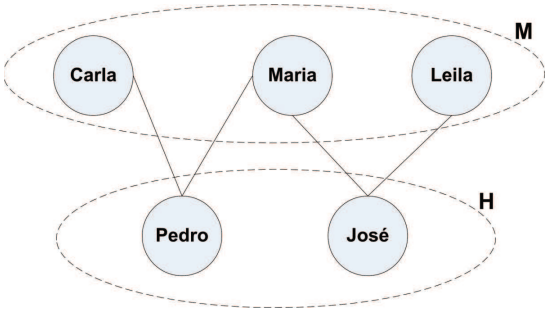


Figura 2.10: Ilustração de um grafo bipartido.

Multigrafo

Um grafo $G = (V,E)$ é denominado de *multigrafo* quando existem múltiplas arestas entre algum par de vértices de G [89].

Considere, como exemplo, o grafo $G = (V, E)$, representado na Figura 2.11. No grafo G, há duas arestas entre os vértices 1 e 2 e entre os vértices 2 e 3.

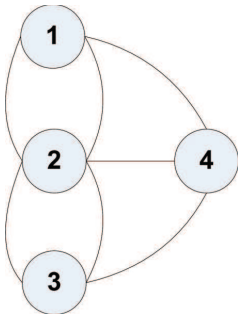


Figura 2.11: Ilustração de um multigrafo.

2.4.4 Subgrafo

Um grafo $H = (V', E')$ é um *subgrafo* de $G = (V, E)$, se $V' \subseteq V$ e $E' \subseteq E$. Um subgrafo pode ser representado na forma $H \subseteq G$. Neste caso, diz-se que G contém H [71].

Seja H um subgrafo de $G = (V, E)$, se $H \subseteq G$ e H contém todas as arestas $(v, u) \in E$, e dado que v, u são vértices de H, então H é um subgrafo induzido de G. Em outras palavras, H é um subgrafo induzido de G, se o mesmo contém todas as arestas que aparecem em G sobre o mesmo conjunto de vértices. Se o conjunto de vértices de H é

um subconjunto S de $V(G)$, então H pode ser escrito como $G[S]$. Logo, diz-se que H é o subgrafo de G induzido por S [91].

Subgrafos podem ser obtidos removendo-se as arestas e vértices de um grafo. Por exemplo, o subgrafo $G - \{v\}$ é obtido ao remover o vértice v e todas as arestas que são incidentes ao mesmo.

Considere, como exemplo, o multigrafo $G = (V, E)$, apresentado na Figura 2.11. Um subgrafo $H = G - \{2\}$, pode ser obtido removendo-se o vértice 2 e todas as suas arestas de G. O subgrafo H é apresentado na Figura 2.12.

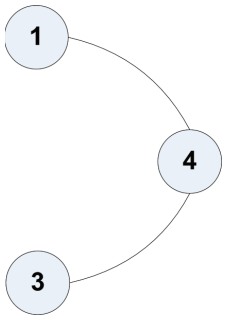


Figura 2.12: Ilustração de um subgrafo induzido.

Clique

Dado um grafo $G = (V, E)$, um *clique* é um grafo $G' = (V', E')$ que é um subgrafo de G e ao mesmo tempo também é um grafo completo [87]. Como exemplo, considere o grafo ilustrado na Figura 2.13.

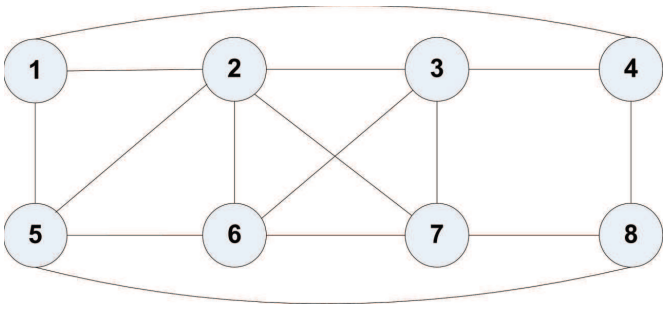


Figura 2.13: Ilustração de um grafo G contendo dois cliques.

No grafo ilustrado na Figura 2.14 há dois cliques. O primeiro clique (grafo à esquerda) é formado pelos vértices $V' = \{1, 2, 5\}$. Por sua vez, o segundo clique (grafo à direita) é formado pelos vértices $V'' = \{2, 3, 6, 7\}$.

2.4.5 Conectividade

Nesta seção são apresentados os conceitos relacionados à conectividade de um grafo.

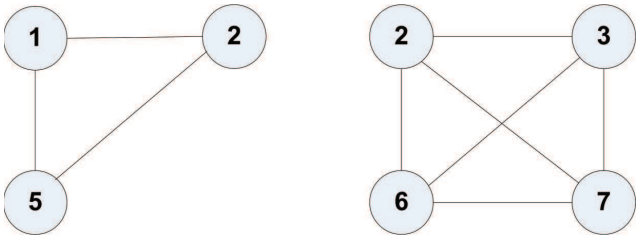


Figura 2.14: Ilustração dos cliques presentes no grafo G.

Caminho

Um caminho é uma sequência de vértices conectados por arestas. De modo geral, o caminho entre v_1 e v_k é formado pela sequência de vértices $s = \{v_1, v_2, \cdots, v_k\}$, tal que $(v_i, v_{i+1}) \in E$, para $i = 1, \cdots, k-1$. Com base na definição anterior, observa-se que pode haver mais de um caminho entre dois vértices que estão conectados [71].

Considere o grafo não-direcionado $G = (V, E)$, onde $V = \{1, 2, 3, 4, 5, 6, 7, 8\}$ e $E = \{(1,2)(1,4), (1,5), (2,3), (2,6), (3,4), (3,7), (4,8), (5,8)\}$, representado na Figura 2.15.

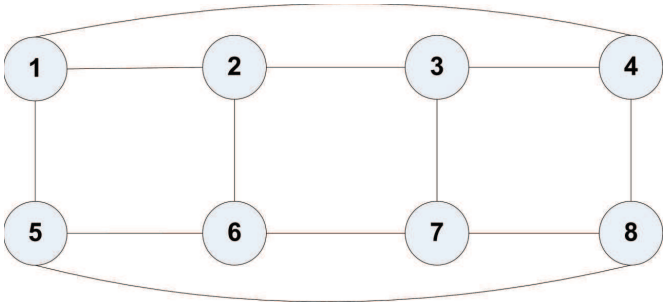


Figura 2.15: Ilustração de um caminho em um grafo simples.

Os seguintes conceitos são apresentados:

- Um *caminho* é dito *simples* se todos os vértices que o compõem são distintos [93]. Por exemplo, na Figura 2.15, um caminho entre os vértices 1 e 7 pode ser formado pela sequência de vértices $s = \{1, 2, 3, 7\}$. Tal sequência é obtida após realizar a travessia sobre as arestas $(1, 2), (2, 3), (3, 7)$ do grafo G;

- O *comprimento de um caminho* é definido como a quantidade de arestas que compõem o caminho. Se existirem mais de um caminho de v_1 a v_2, então o comprimento do caminho de v_1 a v_2 será igual ao menor comprimento dentre todos os caminhos de v_1 a v_2 [93]. Por exemplo, na Figura 2.15, o comprimento do caminho entre os vértices 1 e 7 é $|s| = 3$.

Ciclo

Um ciclo é um caminho simples que começa e termina no mesmo vértice. Isto é, dado um caminho entre v_1 e v_k, formado pela sequência de vértices $s = \{v_1, v_2, \cdots, v_k\}$, onde $v_1 = v_k$ [89].

Por exemplo, na Figura 2.15, um caminho formado pela sequência de vértices $s = \{1, 2, 6, 5, 1\}$, obtido após realizar a travessia sobre as arestas $(1, 2), (2, 3), (3, 7)$, representa um ciclo.

Fecho Transitivo

O *fecho transitivo direto* (FTD) de um vértice v é formado pelo conjunto de todos os vértices que podem ser atingidos por algum caminho iniciando em v [89]. Por exemplo, considere o grafo G ilustrado na Figura 2.16. O FTD do vértice 5 é formado pelo conjunto $\{1, 2, 3, 4, 5, 6\}$. Note que o próprio vértice faz parte de seu FTD, pois ele é alcançável partindo-se dele mesmo.

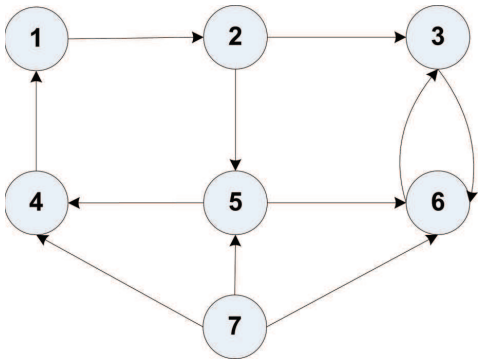


Figura 2.16: Ilustração de um fecho transitivo em um grafo G.

De forma análoga, o *fecho transitivo inverso* (FTI) de um vértice v é formado pelo conjunto de todos os vértices a partir dos quais se pode atingir v por algum caminho [89]. Considere novamente o grafo G, ilustrado na Figura 2.16. O FTI do vértice 5 é formado pelo conjunto $\{1, 2, 4, 5, 7\}$. Note que o próprio vértice faz parte de seu FTI, pois dele se pode alcançar ele mesmo.

Grafo conexo

Um grafo $G = (V, E)$ é *conexo* se existe um caminho entre qualquer par de vértices deste grafo [89].

Como exemplo de um grafo conexo, considere o grafo G ilustrado na Figura 2.17.

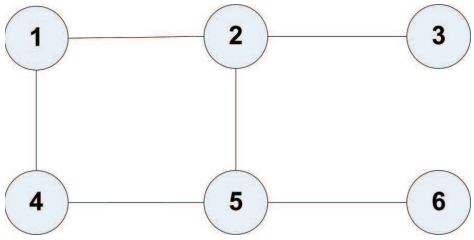


Figura 2.17: Ilustração de um grafo conexo.

Grafo Desconexo

Um grafo $G = (V,E)$ é dito ser *desconexo* se há pelo menos um par de vértices que não está ligado por nenhum caminho [89].

Como exemplo de um grafo desconexo, considere o grafo G ilustrado na Figura 2.18.

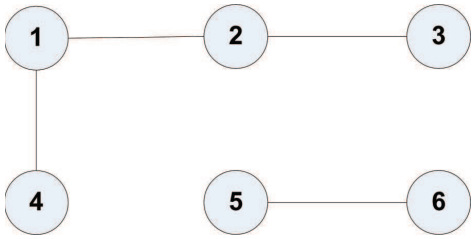


Figura 2.18: Ilustração de um grafo desconexo.

Componente Conexa

Um grafo desconexo $G = (V,E)$ é formado por pelo menos dois subgrafos conexos, disjuntos em relação aos vértices e maximais em relação à inclusão. Cada um destes subgrafos conexos é dito ser uma *componente conexa* de G [89].

O grafo ilustrado na Figura 2.18 possui duas componentes conexas, formadas pelos conjuntos de vértices $V_1 = \{1, 2, 3, 4\}$ e $V_2 = \{5, 6\}$, respectivamente.

Grafo Fortemente Conexo

No caso de grafos orientados, um grafo $G = (V, E)$ é dito ser *fortemente conexo* (F-conexo), se todo par de vértices está ligado por pelo menos um caminho em cada sentido [89]. Ou seja, se cada par de vértices participa de um circuito. Isto significa que cada vértice pode ser alcançável partindo-se de qualquer outro vértice do grafo.

O grafo G, ilustrado na Figura 2.19, é um exemplo de grafo fortemente conexo.

Componente Fortemente Conexa

Um grafo $G = (V,E)$ que não é fortemente conexo é formado por pelo menos dois subgrafos fortemente conexos, disjuntos em relação aos vértices e maximais em relação à inclusão.

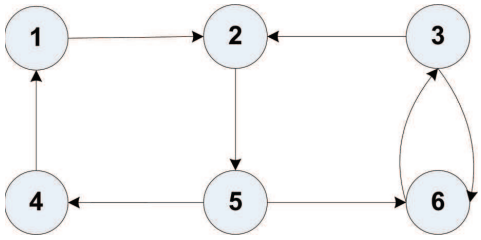


Figura 2.19: Ilustração de um grafo fortemente conexo.

Cada um destes subgrafos é dito ser uma *componente fortemente conexa* de G [89].

Como exemplo de componentes fortemente conexas, considere os subgrafos identificados por S_1, S_2 e S_3, ilustrados na Figura 2.20.

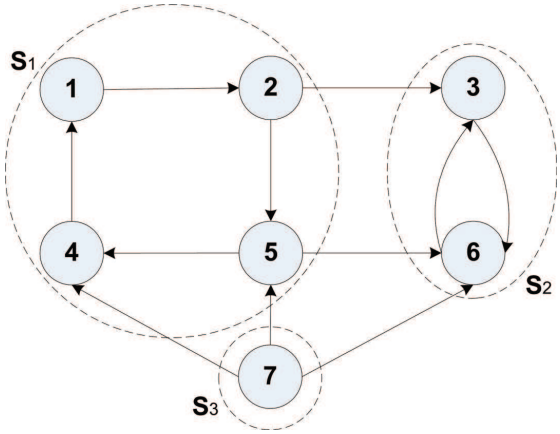


Figura 2.20: Componentes fortemente conexas em um grafo G.

2.4.6 Distância e Breadth-First Search

Conforme discutido anteriormente, a *distância* entre dois vértices em um grafo G pode ser medida com base no menor número de arestas (menor caminho) que conectam os vértices [2].

Em um grafo, a distância entre dois vértices pode ser encontrada utilizando uma técnica denominada de busca em largura (*Breadth-First Search* - BFS) [94]. Em particular, uma busca em largura é iniciada em um vértice inicial, também denominado raiz, e é expandida em níveis sucessivos, conforme descrito a seguir.

1. Partindo de um vértice inicial v, são encontrados todos os vizinhos u de v - isto é, todos os vértices que fazem parte da vizinhança de v. Esses vértices estão na distância 1;

2. Em seguida, todos os vizinhos w de u são encontrados (não devem ser contados os vértices w que são vizinhos de v). Esses vértices estão na distância 2;

3. Então, todos os vizinhos z de w devem ser encontrados (novamente, não devem ser contados os vértices z que foram encontrados nas distâncias 1 e 2). Esses vértices estão na distância 3;

4. Continuando esse processo, a busca será realizada em níveis sucessivos, onde os vértices no nível atual estão distantes em uma unidade em relação aos vértices encontrados no nível imediatamente anterior. Cada nível é constituído de vértices que (i) ainda não foram descobertos nos níveis anteriores e que (ii) possuem uma aresta para algum vértice do nível anterior. Na Figura 2.21, é ilustrado como os níveis são descobertos utilizando uma busca em largura em um grafo.

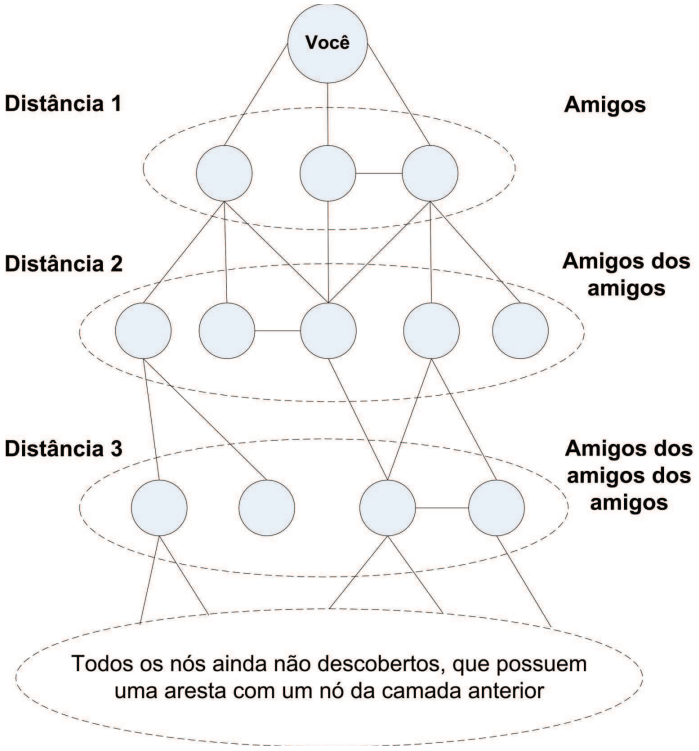


Figura 2.21: Ilustração da busca em largura (Fonte: [2]).

A busca em largura é uma técnica muito utilizada para coletar amostras das estruturas das redes sociais online. Com base nessa técnica, é possível fixar o número máximo de níveis que devem ser expandidos. O grafo resultante desse processo de busca, constitui o que se denomina de conjunto de dados. Um conjunto de dados apresenta tamanho variável, o qual depende da quantidade de usuários conectados, direta e indiretamente, com o vértice inicial e, também, do número de níveis expandidos.

2.5 Considerações Finais

Neste capítulo foi apresentada uma visão geral das teorias e conceitos utilizados para construção da solução descrita nesta tese.

Inicialmente, foi discutido o conceito multidisciplinar de rede social. Assim, procurou-se mostrar que esse conceito é estudado há muito tempo por várias áreas como antropologia, sociologia, comunicação, entre outras. Em particular, os estudos realizados têm como objetivo compreender os processos que fazem com que os indivíduos se relacionem e se organizem em sociedade. Em seguida, mostrou-se que propriedades relacionadas à estrutura das redes sociais são estudadas pela Análise de Redes. Por último, apresentou-se o conceito de redes sociais como um Sistema de Informação na Internet.

Ao longo deste capítulo, também foram discutidos conceitos relacionados a difusão de inovações. Em seguida, foram apresentados os conceitos de influência social e homofilia. Adicionalmente, foi discutido que alguns autores têm encontrado evidências de que esses conceitos podem ser observados nos sistemas de redes sociais, onde eles se confundem, e, portanto, podem ser utilizados em conjunto no processo de difusão de informações.

Em seguida, foram discutidos os modelos de difusão baseados em influência social, comumente utilizados nos trabalhos encontrados na literatura relacionada ao problema de Maximização de Influência baseado em tópicos.

Por fim, foram apresentados os principais conceitos relacionados à Teoria dos Grafos. A Teoria dos Grafos tem sido utilizada como o formalismo para modelar uma rede social e operacionalizar os conceitos relacionados aos elementos das redes sociais.

No próximo capítulo, será apresentada a solução desenvolvida nesta tese para uma variação do problema de Maximização de Influência, denominado Maximização de Influência baseado em Tópicos. Tal solução utiliza vários conceitos das teorias apresentadas neste capítulo.

Capítulo 3

Maximização de Influência Social Ciente de Tópicos

Neste capítulo apresenta-se uma solução para uma variação do problema de Maximização de Influência, denominado de Maximização de Influência baseado em Tópicos. Em particular, a solução proposta utiliza mineração direta sobre o histórico de propagações realizadas pelos usuários em uma rede social, de modo que não é necessário aprender as probabilidades das arestas e executar simulações Monte Carlo para encontrar o conjunto inicial de usuários que maximiza a propagação de informações relacionadas a um tópico específico. Inicialmente, é realizada uma formalização dos conceitos relacionados ao problema abordado. Em seguida, são introduzidos um modelo de Distribuição de Créditos ciente de Tópicos e um conjunto de algoritmos que utilizam os conceitos definidos nesse modelo, como base para a construção de uma solução eficiente e escalável, que possua uma garantia de aproximação em relação à solução ótima.

3.1 Formalização do Problema

Nesta seção são definidos vários conceitos utilizados ao longo do capítulo para resolver o problema abordado neste trabalho. A definição formal desses conceitos é realizada utilizando a Teoria dos Conjuntos [95].

Definição 1 ***(Rede social)*** *Uma rede social é representada como um grafo direcionado* $RS = (U, R)$*, onde* U *é o conjunto de usuários e* R *é o conjunto de relacionamentos existentes entre os usuários. Cada aresta é um par ordenado* (u, v)*, com* $u, v \in U$*, representando uma relação entre* u *e* v*.*

Adicionalmente, com base na direção das arestas do grafo RS, outros dois conjuntos podem ser definidos com base no conceito de vizinhança de um vértice v. O primeiro

conjunto é representado pelos vértices que possuem arestas no sentido de v (i.e. no qual v é o vértice terminal), definido como $N_{in}(v) = \{w \mid (w, v) \in R\}$. Por sua vez, o segundo conjunto é representado pelos vértices diretamente alcançáveis a partir de v. Isto é, os vértices onde as arestas possuem o vértice v como origem, definido como $N_{out}(v) = \{w \mid (v,w) \in R\}$. Por fim, $|N_{in}(v)|$ e $|N_{out}(v)|$ representam a cardinalidade dos respectivos conjuntos.

Definição 2 ***(Registro de Propagações)*** *Um registro de propagações é definido como uma relação Actions(User, Action, Time), na qual tuplas (u,a,t) indicam que o usuário u realizou uma ação a no tempo t. Nesse registro estão armazenadas tuplas para toda ação realizada por todos os usuários de uma rede social.*

Neste trabalho, assume-se que o resultado da operação de projeção sobre a primeira coluna da relação *Actions* está contido no conjunto de usuários U de uma rede social RS. Isto é, os usuários registrados na tabela *Actions* correspondem aos usuários da rede social. Ainda, o conjunto A denota o universo de ações realizadas pelos usuários.

Definição 3 ***(Tópico)*** *Com base na definição encontrada em [20], um tópico T é uma coleção de palavras-chaves relacionadas a um tema específico. Essa coleção é denotada por $T = \{n_1, n_2, \cdots, n_k\}$. Cada um dos n_k temas pode ser uma palavra, uma frase, um meme*[1]*, uma tag, uma URL ou qualquer outro tipo de rótulo que possa ser associado com um vértice.*

Problema 1 *Dados um grafo direcionado representando uma rede social $RS = (U, R)$, um registro de propagações (Actions), um tópico $tp \in T$ e um valor k representando a quantidade de usuários a serem selecionados, o problema de Maximização de Influência baseado em Tópicos, refere-se a encontrar um conjunto $S \subseteq U$ em RS, denominado de conjunto inicial, onde $|S| = k$, de modo que o número esperado de usuários interessados no tópico tp, denotado por $\sigma_m^{tp}(S)$, seja máximo.*

Na próxima seção será apresentada uma visão geral da solução proposta neste trabalho.

3.2 Visão Geral da Solução

A fim de lidar com as características de escala e de dinamicidade que são inerentes a um sistema de redes sociais - mudanças frequentes na estrutura e nos tópicos de interesse dos usuários - e também permitir que a influência social exercida por todos os usuários em cada tópico seja (re)aprendida rapidamente, a solução proposta neste trabalho utiliza

[1]Um meme é uma ideia que se espalha rapidamente de pessoa para pessoa na Web.

uma abordagem para mineração direta do conjunto inicial de usuários, sem a necessidade de executar uma etapa anterior de aprendizagem das probabilidades sobre as arestas do grafo. Além disso, por utilizar um modelo que captura as ações executadas pelos usuários, a solução não necessita de simulações Monte Carlo para selecionar os usuários que irão compor o conjunto inicial. De acordo com a análise realizada sobre os trabalhos relacionados, contida no Capítulo 1, observa-se que as abordagens que utilizam mineração direta do conjunto inicial contemplam várias das características requeridas para resolver o problema abordado, com exceção da influência relacionada aos tópicos de interesse dos usuários. Portanto, a solução proposta neste capítulo consiste de uma extensão dos trabalhos existentes nesta linha de pesquisa, onde é considerada a influência social relacionada aos tópicos de interesse dos usuários.

Especificamente, na abordagem proposta são adicionadas informações sobre os tópicos relacionados às ações presentes no registro de propagações. A partir dessas informações, é possível induzir grafos de propagação baseados em tópicos. Consequentemente, antes de minerar diretamente o conjunto inicial, também podem ser filtradas as tuplas relacionadas a esses tópicos, que contêm os usuários dos respectivos grafos. Uma visão geral da solução proposta é apresentada na Figura 3.1.

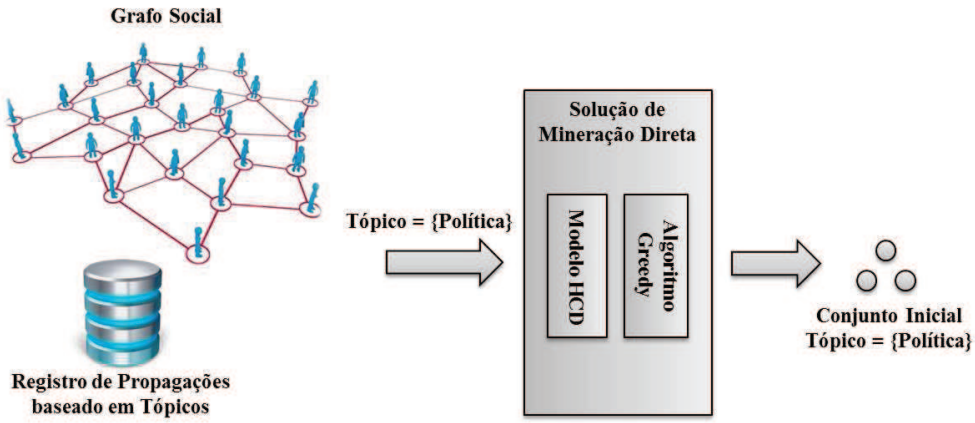


Figura 3.1: Representação da visão geral da solução.

O grafo representando as conexões sociais dos usuários, o registro de propagações baseado em tópicos e um tópico de interesse são fornecidos como entrada para a solução proposta. Especificamente, a solução irá processar todo o registro de propagações com o intuito de descobrir: (i) quais são os tópicos de interesse dos usuários; (ii) quais usuários propagaram informações referentes a cada tópico existente no registro de propagações e; (iii) como esses usuários influenciam seus amigos na propagação de informações relacionadas a esses tópicos. Em seguida, dentre os usuários que realizaram propagações relacionadas ao tópico de interesse, a solução encontrará o conjunto inicial de usuários que maximiza a propagação de informações relacionadas àquele tópico.

É importante destacar que a informação sobre o tópico relacionado a uma ação qualquer pode ser obtida de várias formas, como por exemplo, a partir da utilização de um método de Modelagem de Tópicos, como *Latent Dirichlet Allocation* (LDA) [96]. Especificamente, LDA pode ser utilizado sobre as mensagens compartilhadas entre os usuários em uma rede social, com o objetivo de encontrar os tópicos relacionados a essas mensagens e descobrir quais são os tópicos de interesse dos usuários.

Outra forma de se obter as informações dos tópicos relacionados aos conteúdos compartilhados pelos usuários é explorar a utilização de metadados (por exemplo, *tags*) explicitamente associados aos conteúdos pelos próprios usuários. Neste caso, considere como exemplo o compartilhamento de um vídeo em uma rede social. Antes de compartilhar o vídeo, os usuários associam um conjunto de *tags* ao vídeo, com o objetivo de fornecer um contexto ao vídeo. Esse cenário é comum em alguns tipos de mídias sociais, tais como Digg[2], Delicious[3] e Flickr[4], onde os usuários frequentemente associam *tags* ao conteúdo que está sendo compartilhado. Dessa forma, os usuários podem compartilhar esses conteúdos e as *tags* preferidas com os seus amigos.

Neste trabalho, é assumido que a informação sobre os tópicos relacionados às ações foram fornecidas previamente pelos usuários.

Nas próximas seções são introduzidos os conceitos fundamentais utilizados para construção da solução proposta.

3.3 Modelo de Distribuição de Créditos

O modelo de Distribuição de Créditos (*Credit Distribution model* - CD) foi introduzido originalmente no trabalho de Goyal *et al.* [35]. A ideia básica utilizada neste modelo é que o número esperado de vértices interessados em uma ação, representado por $\sigma_m(S)$, pode ser estimado diretamente a partir das propagações contidas no registro de propagações, conforme definido na Equação (3.1) [35]:

$$\sigma_m(S) = \sum_{u \in U} Pr[path(S, u) = 1] \tag{3.1}$$

Onde:

- $path(S, u)$ é uma variável aleatória que assume o valor 1, se existir um caminho direto do conjunto inicial S para o vértice u. Caso contrário, essa variável assume o valor 0.

[2] `http://digg.com/`
[3] `https://delicious.com/`
[4] `http://www.flickr.com/`

Com base na Equação (3.1), observa-se que o tamanho da propagação esperada, produzida a partir de um conjunto inicial S, pode ser calculado como o somatório de todos os vértices $u \in U$ que foram ativados pelo conjunto S. Essa ativação ocorrerá sempre que existir um caminho direto do conjunto inicial S para o vértice u.

A fim de estimar a probabilidade $Pr[path(S, u) = 1]$, para qualquer conjunto inicial S e vértice u, faz-se necessário que no registro de propagações esteja armazenada uma quantidade enorme de propagações. Desse modo, poder-se-ia explorar todas as combinações possíveis desses dois elementos, onde cada propagação teria como iniciador exatamente o conjunto inicial S que está sendo procurado. Todavia, na prática, tal registro de propagações, com uma combinação claramente exponencial de entradas, não está publicamente disponível para que possa ser utilizado [35].

Assim, para contornar o problema de esparsidade dos dados disponíveis, no modelo de Distribuição de Créditos, a probabilidade $Pr[path(S, u) = 1]$ pode ser estimada utilizando uma perspectiva centrada no usuário. Nesta perspectiva, créditos são atribuídos de forma direta e transitiva aos possíveis influenciadores de um vértice u, toda vez que este vértice executa uma ação.

Considerando o problema abordado neste trabalho, uma importante limitação do modelo de Distribuição de Créditos proposto por Goyal *et al.* [35], é que os créditos são distribuídos sem considerar os possíveis tópicos relacionados às ações realizadas pelos usuários. Consequentemente, assume-se erroneamente que os usuários exercem a mesma influência social sobre qualquer um dos seus amigos, independentemente do tópico considerado.

Nas próximas seções, são introduzidos os conceitos necessários para estender o modelo de Distribuição de Créditos, de modo que o mesmo torne-se ciente dos tópicos relacionados às ações realizadas pelos usuários. Esse novo modelo, denominado de *Topic-Aware Homophily-based Credit Distribution* (HCD), será utilizado como base para estimar o número esperado de usuários interessados em uma ação relacionada a um tópico específico.

3.4 Modelo de Distribuição de Créditos ciente de Tópicos

Nesta seção, descreve-se como estender o Modelo de Distribuição de Créditos para torná-lo ciente de tópicos.

3.4.1 Modelo de Dados

Na Seção 3.1, um registro de propagações foi definido como uma relação *Actions(User, Action, Time)*. Em particular, uma tupla (u, a, t) desta relação indica que o usuário u realizou uma ação a no tempo t. Uma importante limitação desta relação é que a partir das informações disponíveis em qualquer uma de suas tuplas não é possível determinar o tópico relacionado a uma dada ação. Como resultado, quando as tuplas contidas na relação *Actions* são utilizadas no problema de Maximização de Influência baseado em Tópicos, para encontrar os usuários contidos no conjunto inicial S, a influência exercida por esses usuários será a mesma independentemente do tópico considerado.

Na prática, o que ocorre é que cada ação está relacionada a um ou mais tópicos. Por exemplo, considere que uma ação a, realizada por um usuário u, possui a seguinte semântica: "um usuário u avaliou um filme em um dado instante de tempo". Considere ainda que cada filme está relacionado a um ou vários gêneros (por exemplo, gêneros de Ação e Aventura). Dessa forma, a lista de gêneros associados ao filme poderia ser associada ao conjunto de tópicos relacionados a uma ação a.

Desse modo, ao explorar a noção de que toda ação a está relacionada a pelo menos um tópico, faz-se necessário redefinir a relação *Actions* para acomodar essa nova informação. A seguinte definição estende a definição de registro de propagações apresentada anteriormente.

Definição 4 ***(Registro de Propagações baseado em Tópicos)*** *Um registro de propagações baseado em tópicos é definido como uma relação $Actions_T$(User, Action, Topic, Time), a qual contém tuplas (u, a, tp, t). Uma tupla desta relação possui a seguinte semântica: um usuário u realizou uma ação a relacionada ao tópico tp no tempo t.*

Em um registro de propagações baseado em tópicos estão armazenadas tuplas para todas as possíveis ações realizadas por todos os usuários de uma rede social. Neste trabalho, assume-se que o resultado da projeção sobre a primeira coluna da relação $Actions_T$ está contido no conjunto de usuários U de uma rede social RS. Isto é, os usuários registrados na tabela $Actions_T$ correspondem aos usuários da rede social. Ainda, o conjunto A denota o universo de ações que podem ser realizadas pelos usuários. Dessa forma, considera-se que o conjunto A representa todas as ações presentes em $Actions_T$. Por fim, assume-se que uma projeção sobre a terceira coluna da relação $Actions_T$ produzirá como resultado o conjunto de tópicos T, que representa todos os tópicos que podem estar relacionados às ações.

Na Tabela 3.1 está sumarizado um conjunto de tuplas contidas na relação $Actions_T$. Nessa relação, as ações $a \in A = \{Filme_1, Filme_2\}$ são representadas pelos filmes que, por sua vez, estão relacionados a um ou vários tópicos contidos no conjunto de tópicos

Tabela 3.1: Exemplo de um registro de propagação baseado em tópico.

Usuário	Ação	Tópico	Tempo
u_3	$Filme_1$	$tp_1 = \{Drama\}$	t_1
u_2	$Filme_1$	$tp_1 = \{Drama\}$	t_1
u_2	$Filme_2$	$tp_2 = \{Ação\}$, $tp_3 = \{Aventura\}$	t_2
u_1	$Filme_1$	$tp_1 = \{Drama\}$	t_2
u_3	$Filme_2$	$tp_2 = \{Ação\}$, $tp_3 = \{Aventura\}$	t_3
u_5	$Filme_1$	$tp_1 = \{Drama\}$	t_3
u_4	$Filme_1$	$tp_1 = \{Drama\}$	t_4
u_5	$Filme_2$	$tp_2 = \{Ação\}$, $tp_3 = \{Aventura\}$	t_5
u_7	$Filme_1$	$tp_1 = \{Drama\}$	t_5
u_4	$Filme_2$	$tp_2 = \{Ação\}$, $tp_3 = \{Aventura\}$	t_6
u_6	$Filme_2$	$tp_2 = \{Ação\}$, $tp_3 = \{Aventura\}$	t_7

$T = \{tp_1, tp_2, tp_3\}$, onde $tp_1 = \{Drama\}$, $tp_2 = \{Ação\}$ e $tp_3 = \{Aventura\}$. Observe que uma ação pode estar relacionada a um ou vários tópicos. Por exemplo, $Filme_1$ está relacionado aos tópicos tp_2 e tp_3. Além disso, um usuário (por exemplo, u_2 e u_3) pode realizar várias ações e essas podem estar relacionadas a tópicos diferentes.

Na seguinte definição é formalizado como uma ação relacionada a um tópico é propagada entre os usuários de uma rede social.

Definição 5 ***(Propagação de ação)*** *Uma ação $a \in A$, relacionada a um tópico $tp \in T$, é propagada de um usuário u_i para outro usuário u_j $\iff$ $(u_i, u_j) \in R$ e $\exists (u_i, a, tp, t_i)$, $(u_j, a, tp, t_j) \in Actions_T$, com $t_i < t_j$.*

Observe que, de acordo com a Definição 5, para que uma ação seja propagada entre dois usuários u_i e u_j, na rede social RS, deve existir uma aresta relacionando u_i e u_j. Além disso, ambos os usuários devem ter realizado a mesma ação, um necessariamente antes do outro, com base na restrição temporal utilizada. Tal fato permite que seja construído um grafo de propagação baseado em tópico, utilizando como base a próxima definição.

Definição 6 ***(Grafo de Propagação baseado em Tópico)*** *Para cada ação $a \in A$, relacionada a um tópico $tp \in T$, um grafo de propagação baseado em tópico é definido como $PG^{tp}(a) = (U^{tp}(a), R^{tp}(a))$, onde $U^{tp}(a) = \{v \mid \exists\, (u, a, tp, t) \in Actions_T\}$. Ainda, existe uma aresta direcionada $u_i \xrightarrow{\Delta t} u_j$ toda vez que uma ação a, relacionada a um tópico tp, for propagada de u_i para u_j, com (u_i, a, tp, t_i), $(u_j, a, tp, t_j) \in Actions_T$, onde $\Delta t = t_j - t_i$.*

Um grafo de propagação baseado em tópico consiste nos usuários que realizaram uma ação relacionada a um tópico específico, onde as arestas que conectam os vértices estão orientadas no sentido da propagação. Observe que cada grafo de propagação baseado em tópico é um grafo acíclico dirigido (DAG), o qual apresenta as seguintes propriedades: (i) cada vértice pode ter mais de um pai; (ii) as arestas são direcionadas no sentido da propagação; (iii) ciclos são impossíveis devido à restrição temporal, que é a base da

definição da propagação e; (iv) o grafo de propagação resultante pode ter componentes desconectadas.

Quando observada sob outra perspectiva, a propagação de uma ação relacionada a um tópico $tp \in T$ no grafo $PG^{tp}(a)$, é na verdade um fluxo que se espalha de forma transitiva entre os vértices, cujo o sentido é determinado pelos vértices que realizaram aquela ação. Por sua vez, o registro de propagações baseado em tópicos $Actions_T$*(User, Action, Topic, Time)* pode ser visto como uma coleção de grafos de propagação baseado em tópico.

Na Figura 3.2, é ilustrado o conceito de grafo de propagação baseado em tópico. Na Figura 3.2(a), é ilustrado o grafo RS construído a partir das informações contidas na Tabela 3.1. Por sua vez, nas Figuras 3.2(b) e 3.2(c), são ilustrados os grafos de propagação baseado em tópico referentes aos filmes relacionados aos tópicos $tp_2 = \{Ação\}$ e $tp_1 = \{Drama\}$, respectivamente.

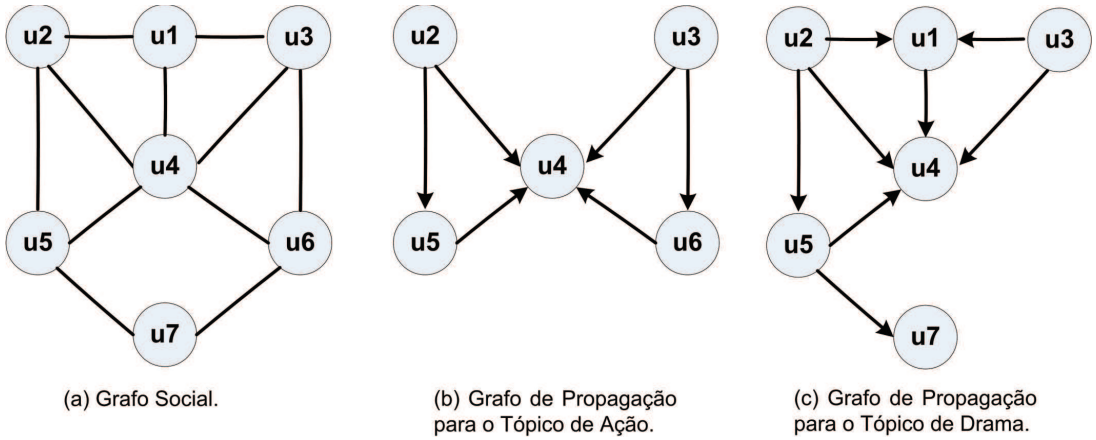


Figura 3.2: Ilustração de como induzir grafos de propagação baseado em tópico.

Nas próximas seções é descrito como calcular os créditos do modelo de Distribuição de Créditos ciente de Tópicos.

3.4.2 Distribuição de Créditos Diretos utilizando Homofilia

A notação utilizada nas próximas seções está sumarizada na Tabela 3.2.

Quando um usuário realiza uma ação a, são distribuídos créditos diretos aos seus amigos que realizaram a mesma ação, por eles terem influenciado o usuário naquela ação. Como restrição, tem-se que o somatório dos créditos diretos que um usuário distribui aos seus amigos, por o influenciarem em uma dada ação, não pode ser maior do que 1. Em particular, a função de distribuição de créditos diretos é dada pela Equação (3.2), a qual foi apresentada originalmente no trabalho de Goyal *et al.* [35], e estendida neste trabalho, de modo a distribuir créditos diretos de acordo com um determinado tópico de interesse e também com base na similaridade entre os usuários.

Tabela 3.2: Notação utilizada para definição do modelo HCD.

$\mathcal{A}_u^{tp}$	Número de ações relacionadas a um tópico $tp \in T$ que foram realizadas por um usuário u.
$\mathcal{NA}_u^{tp}$	Número de ações relacionadas a um tópico $tp \in T$ que foram realizadas por um usuário u, mas que não foram iniciadas pelo mesmo.
$N_{in}^{tp}(u,a)$	Vizinhos do usuário u (i.e. potenciais influenciadores) que executaram uma ação a relacionada ao tópico $tp \in T$ antes de u.
$\gamma_{v,u}^{tp}(a)$	Créditos de influência direta dados a v por ter influenciado u em uma ação a relacionada ao tópico $tp \in T$.
$\Gamma_{v,u}^{tp}(a)$	Total de Créditos dados a v por ter influenciado u em uma ação a relacionada ao tópico $tp \in T$.
$\kappa_{v,u}^{tp}$	Total de créditos dados a v por ter influenciado u em todas as ações relacionadas ao tópico $tp \in T$.
$\tau_{v,u}^{tp}$	Tempo médio para que ações relacionadas ao tópico $tp \in T$ sejam propagadas do usuário u para o usuário v.

$$\gamma_{v,u}^{tp}(a) = \frac{1}{N_{in}^{tp}(u,a)} \cdot \frac{\left[infl_u^{tp} \cdot e^{\left(-\frac{\Delta_{u,v}^{tp}(a)}{\tau_{v,u}^{tp}} \right)} + \left(\sum_{k=1}^{|H|} sim(H_{k_v}, H_{k_u}) \right) \right]}{1 + |H|} \tag{3.2}$$

O termo $infl_u^{tp}$ representa a influenciabilidade do usuário u na realização de ações relacionadas a um tópico $tp \in T$. Especificamente, a influenciabilidade do usuário u pode ser calculada com base na razão entre o número total de ações relacionadas ao tópico tp realizadas pelo usuário, onde ele não foi o iniciador dessas ações, sobre o número total de ações relacionadas ao tópico tp que foram realizadas pelo usuário. A influenciabilidade do usuário é calculada com base na Equação (3.3):

$$infl_u^{tp} = \frac{\mathcal{NA}_u^{tp}}{\mathcal{A}_u^{tp}} \tag{3.3}$$

Com base na Equação (3.3), observa-se que a influenciabilidade do usuário u depende diretamente da quantidade total de ações relacionadas ao tópico tp que ele não iniciou. Logo, quanto mais ações relacionadas ao tópico tp forem realizadas pelo usuário, onde ele não foi o iniciador dessas ações, mais influenciável será o usuário. De forma contrária, o usuário será menos influenciável caso ele seja o iniciador da maior parte das ações relacionadas a um tópico tp.

Adicionalmente, a influenciabilidade de um usuário u também depende do tempo em que o usuário executou a ação. Conforme pode ser observado na Equação (3.2), foi adicionada uma dependência temporal que faz com que a influenciabilidade do usuário u diminua exponencialmente no decorrer do tempo. A ideia básica é que quanto maior o intervalo de tempo observado para que u execute a mesma ação realizada por v, menor será a influenciabilidade de u.

Para calcular o fator de decaimento, pode-se calcular o tempo médio para que as ações

a relacionadas a um tópico $tp \in T$ sejam propagadas do usuário v para o usuário u. Essa informação pode ser calculada utilizando um subconjunto das tuplas contidas no registro de propagações baseado em tópicos.

O outro termo necessário para calcular o fator de decaimento é a função delta, definida na Equação (3.4):

$$\Delta_{u,v}^{tp}(a) = t(u,a) - t(v,a) \tag{3.4}$$

Especificamente, na Equação (3.4) é calculada a diferença entre os tempos de execução de uma ação a, relacionada a um tópico tp, por parte dos usuários u e v. O tempo de execução de uma ação relacionada a um tópico tp, para os usuários u e v, é calculado a partir das funções $t(u,a)$ e $t(v,a)$, respectivamente.

Além de introduzir uma dependência em relação ao tópico na função de distribuição de créditos diretos, neste trabalho, também foi adicionada uma dependência em relação às características presentes no arcabouço teórico de Homofilia.

Do arcabouço teórico de homofilia, sabe-se que os usuários tendem a se relacionar com pessoas que possuem características similares. Por exemplo, as pessoas tendem a formar laços de amizade com outras pessoas que sejam do mesmo gênero, tenham a mesma idade, sejam da mesma família, trabalhem ou estudem no mesmo local, moram na mesma cidade, sejam da mesma raça, sejam da mesma classe social, etc.

Dessa forma, a ideia chave é que a influenciabilidade do usuário possa aumentar ou diminuir, de acordo com a similaridade existente entre o usuário e seus amigos. Por exemplo, considere que os usuários u e v sejam amigos e que o usuário z é um amigo em comum. Ainda, considere que u e v possuem várias características de homofilia em comum (por exemplo, possuem a mesma idade e gênero e trabalham no mesmo local). Assim, supõem-se que, quando v realiza uma ação e compartilha a mesma com sua rede de amigos, o usuário u estará mais propenso a realizar a mesma ação do que o usuário z.

Especificamente, a dependência em relação às características de homofilia foi introduzida na Equação (3.2), a partir do somatório de seus fatores, dado por $\sum_{k=1}^{|H|} sim(H_{k_v}, H_{k_u})$. Adicionalmente, o fator $1 + |H|$ é utilizado para garantir que o somatório entre a influenciabilidade do usuário e os fatores relacionados à homofilia varie dentro do intervalo $[0, 1]$.

Por fim, para garantir que o somatório dos créditos diretos não seja maior do que 1, esses créditos são normalizados entre os vizinhos $v \in N_{in}^{tp}(u,a)$ do usuário u, que realizaram a ação antes dele. Isto é, entre os potenciais influenciadores do usuário u.

Na próxima seção é descrito como calcular a similaridade entre os usuários para as informações relacionadas a homofilia.

Cálculo de Similaridade

Neste trabalho, dentre os conceitos existentes no arcabouço teórico de homofilia, foram utilizadas apenas algumas informações relacionadas ao perfil público do usuários (gênero e a idade), para a definição da função de distribuição de créditos diretos. De modo a abstrair a introdução de novas variáveis relacionadas a homofilia à função de similaridade, optou-se por acomodar essas variáveis dentro de um conjunto H.

Dessa forma, H_k representa o k-ésimo elemento do conjunto de fatores relacionados a homofilia, quais sejam: $H = \{age, gender\}$. Para cada elemento H_k do conjunto H, é realizado o calculo da similaridade entre os usuários v e u para esse elemento, representado por $sim(H_{k_v}, H_{k_u})$. A seguir são apresentadas as equações utilizadas para calcular a similaridade de idade e gênero.

Similaridade de Idade

O cálculo da similaridade entre as idades de dois usuários pode ser realizada com base na comparação dessas idades, utilizando-se como referência as faixas etárias nas quais os usuários estão inseridos. Especificamente, uma faixa etária pode ser representada através de um intervalo $[min_{age}, max_{age}]$, que representam a idade mínima e a idade máxima, respectivamente. Neste trabalho, a Equação (3.5) é utilizada para calcular a similaridade entre as idades de dois usuários:

$$sim(age_v, age_u) = \alpha^{rd} \tag{3.5}$$

Onde:

- age_{user} é uma função indicadora que retorna um número inteiro representando a idade do usuário;
- α é um fator de atenuação. Por exemplo, $\alpha = 0{,}5$;
- rd é o valor absoluto da diferença entre a ordem de classificação das faixas etárias correspondentes às idades dos usuários v e u, respectivamente.

A ideia chave é que o valor da similaridade entre as idades dos usuários sofre uma atenuação à medida em que as faixas etárias que as representam se distanciam uma da outra. Um exemplo prático de utilização da Equação (3.5) é descrito a seguir.

Considere que os usuários João (18 anos), Maria (18 anos), José (25 anos) e Pedro (35 anos) são amigos entre si. Utilizando $\alpha = 0{,}5$ e os valores sumarizados na Tabela 3.3, como referência para determinar a faixa etária correspondente à idade desses usuários,

Tabela 3.3: Faixa etária utilizada para o cálculo de similaridade de idade.

Ordem de Classificação	Idade	Faixa Etária
1	1	(<18)
2	18	(18-24)
3	25	(25-34)
4	35	(35-44)
5	45	(45-49)
6	50	(50-55)
7	56	(>56)

pode-se facilmente calcular o valor da similaridade de idade dos usuários João e Maria, bem como, de Pedro e João.

Com base nos valores contidos na Tabela 3.3, observa-se que João e Maria estão classificados na mesma faixa etária. Desse modo, tem-se que o valor absoluto da diferença entre a ordem de classificação das faixas etárias de ambos os usuários é igual a zero ($rd = 0$). Logo, utilizando a Equação (3.5), o valor da similaridade entre as idades de João e Maria, representada por $sim(age_{Joao}, age_{Maria})$, é igual a 1,0. De forma análoga, tem-se que a similaridade entre as idades de Pedro e João, representada por $sim(age_{Pedro}, age_{Joao})$, é igual 0,25, pois a idade de Pedro está classificada duas faixas etárias acima ($rd = 2$) daquela na qual a idade de João está classificada.

Similaridade de Gênero

O gênero de um usuário pode assumir os valores masculino ou feminino, o qual é representado formalmente por $gender_{user} \in \{M, F\}$. Para calcular o nível de similaridade entre os gêneros dos usuários u e v, é utilizada a Equação (3.6):

$$sim(gender_v, gender_u) = \begin{cases} 1{,}0 & \text{se } gender_v = gender_u; \\ 0{,}5 & \text{se } gender_v = \emptyset \ \vee gender_u = \emptyset; \\ 0{,}25 & \text{se } gender_v \neq gender_u. \end{cases} \tag{3.6}$$

Observe que os valores anteriores também podem ser determinados por meio de uma função análoga àquela definida na Equação (3.5). Além disso, o valor 0,5 é atribuído sempre que a informação sobre o gênero de pelo menos um dos usuários não estiver disponível.

3.4.3 Distribuição de Créditos Totais

Além de créditos diretos, também são distribuídos créditos de forma transitiva entre os usuários que realizaram uma ação relacionada ao tópico $tp \in T$. Dessa forma, quando um usuário v, que é amigo direto de u, recebe créditos diretos por ter influenciado u na realização de uma ação relacionada ao tópico tp, ele também deve distribuir esses créditos entre os seus amigos que realizaram aquela ação.

Na Equação (3.7) é definida a função de distribuição de créditos total.

$$\Gamma^{tp}_{v,u}(a) = \sum_{w \in N^{tp}_{in}(u,a)} \Gamma^{tp}_{v,w}(a) \cdot \gamma^{tp}_{w,u}(a) \tag{3.7}$$

Na Equação (3.7), os créditos totais são distribuídos de forma recursiva pelo usuário u a todos os usuários contidos no grafo de propagação baseado em tópico, correspondente à ação a. A base da recursão é $\Gamma^{tp}_{v,v}(a) = 1$.

Por sua vez, na Equação (3.8), é definido como calcular o crédito total recebido pelo conjunto inicial $S \subseteq U^{tp}(a)$, devido aos usuários pertencentes ao conjunto terem influenciado o usuário u na realização de uma ação a relacionada ao tópico tp.

$$\Gamma^{tp}_{S,u}(a) = \begin{cases} 1 & \text{se } v \in S; \\ \displaystyle\sum_{w \in N^{tp}_{in}(u,a)} \Gamma^{tp}_{S,w}(a) \cdot \gamma^{tp}_{w,u}(a) & \text{caso contrário.} \end{cases} \tag{3.8}$$

Agregação dos Créditos Totais sobre todas as Ações de um Tópico

De modo geral, o crédito total recebido pelo usuário v, devido a este ter influenciado o usuário u na realização de todas as ações $a \in A$, que estão relacionadas ao tópico tp, pode ser calculado utilizando a Equação (3.9):

$$\kappa^{tp}_{v,u} = \frac{1}{A^{tp}_u} \sum_{a \in A} \Gamma^{tp}_{v,u}(a) \tag{3.9}$$

De forma similar, o crédito total recebido por um conjunto inicial $S \subseteq U^{tp}(a)$, devido aos usuários pertencentes ao conjunto terem influenciado o usuário u na realização de todas as ações $a \in A$, relacionadas ao tópico $tp \in T$, pode ser calculado utilizando a Equação (3.10):

$$\kappa^{tp}_{S,u} = \frac{1}{A^{tp}_u} \sum_{a \in A} \Gamma^{tp}_{S,u}(a) \tag{3.10}$$

Por fim, o crédito total recebido por um conjunto inicial $S \subseteq U^{tp}(a)$, devido aos

usuários pertencentes ao conjunto terem influenciado todos os usuários do grafo G na realização de ações relacionadas ao tópico tp (i.e. $U^{tp} \subseteq U$), pode ser calculado utilizando a Equação (3.11):

$$\sigma^{tp}_{hcd} = \sum_{u \in U^{tp}} \kappa^{tp}_{S,u} \tag{3.11}$$

A Equação (3.11) corresponde à função objetivo que deve ser maximizada para resolver o problema abordado neste trabalho.

3.4.4 Garantia de Aproximação

No trabalho de Goyal *et al.* [35] foi demonstrado que o problema de Maximização de Influência utilizando o modelo de Distribuição de Créditos é NP-Difícil. Além disso, os autores também demonstraram que a função objetivo, definida sobre todas as ações, independentemente do tópico, é monótona e submodular.

Conforme argumentado anteriormente, o modelo HCD é um caso particular do modelo CD, onde todas as ações pertencentes a um mesmo tópico estão agregadas. Dessa forma, considerando o conjunto de tópicos $T = \{tp_1, tp_2, tp_3\}$, se apenas as ações relacionadas a um tópico específico forem utilizadas, por exemplo, $tp_1 = \{Drama\}$, então, o modelo HCD se comportará de modo similar ao modelo CD para o tópico escolhido e apresentará as mesmas propriedades deste modelo. Além disso, como a função objetivo, definida na Equação (3.11), é um caso particular daquela definida sobre todas as ações, então, esta função também é monótona e submodular.

Com base no resultado anterior, conforme apontado por Kempe *et al.* [34], pode-se, então, construir uma solução baseada no algoritmo *Greedy* [37], para encontrar o conjunto inicial S, que oferece uma garantia de aproximação de $(1-1/e)$ em relação à solução ótima.

Na próxima seção são descritos os passos necessários para resolver o problema de Maximização de Influência baseado em Tópicos.

3.5 Descrição da Solução

Os passos descritos a seguir utilizam o registro de propagações baseado em tópicos e o modelo de Distribuição de Créditos ciente de tópicos, descritos nas seções anteriores, como base para a construção de uma solução de mineração direta de dados para o problema de Maximização de Influência baseado em Tópicos.

1. Inicialmente, processe previamente todo o registro de propagações baseado em tópicos ($Actions_T$). Durante este processamento, para todas as combinações de vértices

u e v e ações a relacionadas a cada um dos tópicos $tp \in T$, calcule e armazene os valores de $\tau_{v,u}^{tp}$ e $infl^{tp}(u)$;

2. Processe novamente todo o registro de propagações $Actions_T$ (Algoritmo 1). Dessa vez, realize os seguintes passos:

   (a) Calcule os créditos diretos $\gamma_{v,u}^{tp}(a)$, para todas as combinações de vértices u e v e ações relacionadas a tópicos $tp \in T$, utilizando como base os valores de $\tau_{v,u}^{tp}$ e $infl^{tp}(u)$, calculados no passo anterior;

   (b) Calcule os créditos totais $\Gamma_{v,u}^{tp}(a)$, para todas as combinações de vértices u e v e ações relacionadas a tópicos $tp \in T$, e armazene-os na estrutura de dados UC (*User Credits*).

3. Utilize a estrutura de dados UC como entrada para o Algoritmo 2 (algoritmo *Greedy* com implementação CELF [37]), em conjunto com os parâmetros k e $tp \in T$, que representam o tamanho do conjunto inicial S que deve ser encontrado e o tópico de interesse, respectivamente. Em cada iteração do algoritmo, adicione ao conjunto inicial S o vértice x que provê o máximo ganho marginal em relação ao conjunto inicial da iteração anterior;

4. Utilize o Algoritmo 3 para calcular o ganho marginal provido por um vértice x em relação ao conjunto inicial S atual;

5. Após adicionar um vértice x ao conjunto inicial S, utilize o Algoritmo 4 para atualizar de modo incremental as estruturas de dados UC e SC.

3.6 Algoritmos

Nesta seção são descritos os algoritmos utilizados para encontrar o conjunto inicial que soluciona o problema abordado neste trabalho.

Os algoritmos descritos nesta seção utilizam duas estruturas de dados para armazenar as informações sobre os créditos distribuídos no modelo HCD. A estrutura de dados UC mantém informações relacionadas aos créditos totais dados a um vértice v, por este ter influenciado qualquer outro vértice u em todas as ações relacionadas a um tópico tp. Especificamente, cada entrada $UC[v][u][tp][a]$ corresponde ao termo $\Gamma_{v,u}^{tp}(a)$, definido sobre o conjunto $U^{tp} - S$, onde U^{tp} corresponde aos vértices que realizaram ações relacionadas ao tópico tp e S corresponde ao conjunto inicial. Adicionalmente, a estrutura de dados SC (*Set Credits*) mantém informações relacionadas aos créditos totais dados ao conjunto

inicial S pelo vértice x por tê-lo influenciado nas ações relacionadas ao tópico tp. Portanto, cada entrada $SC[x][tp][a]$ corresponde ao cálculo de $\Gamma^{tp}_{S,x}(a)$.

Algoritmo 1: Scan

Input : RS, $Actions_T$, λ, τ^{tp}, $infl^{tp}$
Output: UC

```
1  UC ← ∅;
2  foreach action a in Actions_T do
3      current_table ← ∅;
4      foreach tuple ⟨u, a, T, t_u⟩ in chronological order do
5          foreach topic tp ∈ T do
6              Parents^tp(u) ← ∅;
7              A_u^tp ← A_u^tp + 1;
8              UC[∗][u][tp][a] ← 0;
9              while ∃v : (v,u) ∈ RS, v ∈ current_table do
10                 Parents^tp(u) ← Parents^tp(u) ∪ {v};
11                 foreach v ∈ Parents^tp(u) do
12                     compute γ^tp_{v,u}(a) using Equation 3.2;
13                     if γ^tp_{v,u}(a) ≥ λ then
14                         UC[v][u][tp][a] ← UC[v][u][tp][a] + γ^tp_{v,u}(a);
15                         foreach w such that UC[w][v][tp][a] · γ^tp_{v,u}(a) ≥ λ do
16                             UC[w][u][tp][a] ← UC[w][u][tp][a] + UC[w][v][tp][a] · γ^tp_{v,u}(a);
17                         end
18                     end
19                 end
20             end
21         end
22     end
23     current_table ← current_table ∪ {u}
24 end
```

O Algoritmo 1 é utilizado para calcular os créditos totais para todas as combinações de usuários u e v e ações a relacionadas a tópicos $tp \in T$. Este algoritmo recebe como entradas um grafo representando uma rede social $RS = (U, R)$, um registro de propagações baseado em tópicos ($Actions_T$) e um limiar λ. Como pré-condições para execução deste algoritmo, tem-se que: (1) as informações sobre os gêneros e idades de cada usuário devem estar disponíveis no grafo RS e; (2) as tuplas do registro de propagações baseado em tópicos devem estar ordenadas, primeiro pela ação (i.e. cada ação corresponde a uma propagação

contida em $Actions_T$) e depois pelo tempo, em ordem cronológica. O Algoritmo 1 é descrito a seguir.

A execução do algoritmo é iniciada com o processamento de cada ação contida no registro de propagações baseado em tópicos. Para cada ação são processadas suas tuplas de forma sequencial. Durante esse processamento, é armazenada a quantidade total de ações relacionadas ao tópico $tp \in T$, que foram realizadas pelo usuário u ($\mathcal{A}_u^{tp}$). Todos os amigos de u, isto é, usuários $v \in N_{in}^{tp}(u,a) \wedge v \in current\_table$, que executaram a ação atual, relacionada ao tópico tp, são armazenados em $Parents^{tp}(u)$. Em seguida, para todos os usuários em $Parents^{tp}(u)$, são distribuídos créditos diretos (Equação 3.2) e também créditos totais (Equação 3.7). O parâmetro λ especifica um limiar utilizado para diminuir a quantidade de memória utilizada, uma vez que apenas são armazenadas entradas em UC para os créditos totais que estejam acima desse limiar. Em cada iteração do *loop* principal, é mantida uma lista ($current\_table$) dos usuários que realizaram a ação corrente e que foram encontrados até o momento nas tuplas processadas. A execução do algoritmo é finalizada quando todas as ações forem processadas.

É importante destacar que as informações contidas na estrutura de dados SC não são atualizadas durante a execução do Algoritmo 1.

Por sua vez, o Algoritmo 2, que implementa o algoritmo *Greedy* com a otimização CELF [37], é utilizado para encontrar o conjunto inicial S que maximiza a propagação de ações relacionadas ao tópico $tp \in T$. Especificamente, o Algoritmo 2 recebe como entradas a estrutura de dados UC, computada anteriormente, um parâmetro k, representando o tamanho do conjunto S, e um tópico $tp \in T$. Neste algoritmo é mantida uma Fila Q, na qual são armazenadas internamente entradas no formato $\langle x, mg, it\rangle$, onde mg denota o ganho marginal provido pelo vértice x em relação ao conjunto inicial S, na iteração it. Essas entradas são sempre mantidas organizadas em ordem decrescente, de acordo com o ganho marginal provido por cada vértice x. O Algoritmo é descrito a seguir.

Inicialmente, é encontrado o conjunto $U^{tp} \subseteq U$, o qual contém os vértices que realizaram ações relacionadas ao tópico $tp \in T$. Em seguida, para cada um dos vértices $u \in U^{tp}$, é calculado o ganho marginal provido pelo vértice u e, posteriormente, este vértice é inserido na fila Q. O restante do algoritmo é executado como segue.

Em cada iteração do *loop while*, o primeiro elemento x da fila Q é analisado e uma das seguintes decisões é tomada:

- Se x já foi analisado na iteração atual (i.e. $x.it = |S|$), então x é selecionado e depois inserido no conjunto S. Em seguida, o Algoritmo 4 é executado para atualizar de modo incremental as estruturas de dados UC e SC;
- Caso x tenha sido analisado apenas na iteração anterior (i.e. $x.it < |S|$), então, o

Algoritmo 3 é executado para atualizar o ganho marginal provido pelo vértice x em relação ao conjunto inicial S. Após essa atualização, o vértice x é reinserido na fila Q e a mesma é reordenada.

A execução do algoritmo será finalizada quando o número requerido de vértices contidos em S for atingido (i.e. $|S| = k$).

Algoritmo 2: Greedy com CELF

Input : UC, k, $tp \in T$
Output: Seed set S

```
1  SC ← ∅;
2  S ← ∅;
3  Q ← ∅;
4  foreach u such that ∃u : A_u^tp > 0 do
5      U^tp ← U^tp ∪ {u};
6  end
7  foreach u ∈ U^tp do
8      x.node ← u;
9      x.mg ← computeMG(x, UC, SC);
10     x.it ← 0;
11     add x to Q;
12 end
13 while |S| < k do
14     x ← pop(Q);
15     if x.it = |S| then
16         S ← S ∪ {x};
17         update(x, UC, SC);
18     end
19     else
20         x.mg ← computeMG(x, UC, SC);
21         x.it ← |S|;
22         Reinsert x into Q and heapify
23     end
24 end
```

O Algoritmo 3 é utilizado para calcular o ganho marginal que será provido pelo vértice x ao conjunto inicial S, caso o mesmo seja adicionado ao conjunto. De forma contrária às soluções que utilizam Simulações Monte Carlo, neste trabalho, o cálculo do ganho marginal é realizado de modo incremental, utilizando como base o próprio registro de

propagações baseado em tópicos.

Algoritmo 3: computeMG

Input : x, UC, SC, $tp \in T$

Output: mg

1 $mg \leftarrow 0$;
2 **foreach** *action a such that* $\exists u : UC[x][u][tp][a] > 0$ **do**
3 $\quad mg_a^{tp} \leftarrow 1/\mathcal{A}_x^{tp}$;
4 $\quad$ **foreach** *u such that* $UC[x][u][tp][a] > 0$ **do**
5 $\quad\quad mg_a^{tp} \leftarrow mg_a^{tp} + UC[x][u][tp][a]/\mathcal{A}_u^{tp}$;
6 $\quad$ **end**
7 $\quad mg \leftarrow mg + mg_a^{tp}(1 - SC[x][tp][a])$;
8 **end**

Por fim, o Algoritmo 4 é utilizado para atualizar de modo incremental as estruturas de dados UC (linha 4) e SC (linha 6), respectivamente.

Algoritmo 4: update

Input : x, UC, SC, $tp \in T$

1 **foreach** *action a such that* $\exists u : UC[x][u][tp][a] > 0$ **do**
2 $\quad$ **foreach** *u such that* $UC[x][u][tp][a] > 0$ **do**
3 $\quad\quad$ **foreach** *v such that* $UC[v][x][tp][a] > 0$ **do**
4 $\quad\quad\quad UC[v][u][tp][a] \leftarrow UC[v][u][tp][a] - UC[v][x][tp][a] \cdot UC[x][u][tp][a]$;
5 $\quad\quad$ **end**
6 $\quad\quad SC[u][tp][a] \leftarrow SC[u][tp][a] + UC[x][u][tp][a] \cdot (1 - SC[x][tp][a])$;
7 $\quad$ **end**
8 **end**

A prova do teorema utilizado como base para calcular o ganho marginal provido por um vértice x e também para atualizar as estruturas de dados UC e SC, de modo incremental, está fora do escopo deste trabalho. A mesma pode ser encontrada em [35].

3.7 Considerações Finais

Neste capítulo foi apresentada uma solução para o problema de Maximização de Influência baseado em Tópicos. Em particular, para solucionar tal problema de forma eficiente, foram estendidos vários conceitos do Modelo de Distribuição de Créditos. Neste novo modelo, denominado Modelo de Distribuição de Créditos ciente de Tópicos, foi introduzido o conceito de registro de propagações baseado em tópicos e foram redefinidas as equações utilizadas para calcular os créditos diretos e totais dados a um conjunto de usuários, devido aos mesmos terem influenciado seus amigos em várias ações relacionadas a um

dado tópico. Além disso, na equação utilizada para distribuir os créditos diretos, foram adicionadas informações relacionadas ao arcabouço teórico de homofilia.

Em seguida, foi apresentada uma solução de mineração direta de dados, que utiliza as propagações existentes no registro de propagações baseado em tópicos para alimentar o modelo HCD. Tal solução, não necessita aprender as probabilidades das arestas e, tampouco, em cada iteração do algoritmo *Greedy*, na etapa de seleção dos usuários, executar simulações MC para calcular o ganho marginal provido por um usuário em relação ao conjunto inicial. Por fim, foram descritos os algoritmos utilizados para processar de forma eficiente o registro de propagações baseado em tópicos e encontrar rapidamente o conjunto inicial procurado no problema abordado neste trabalho.

Diferentemente das soluções existentes para o problema abordado, a solução apresentada neste trabalho é escalável. Isto é, independentemente da escala de uma rede social, é possível encontrar, rapidamente e eficientemente, o conjunto inicial que maximiza a propagação de informações relacionadas a um tópico de interesse. Outro diferencial, é que a solução apresentada também oferece uma garantia de aproximação de $(1 - 1/e)$ em relação à solução ótima.

No Capítulo 4 será descrita a avaliação experimental realizada sobre a solução apresentada.

Capítulo 4

Avaliação Experimental

Neste capítulo descreve-se a realização de um projeto experimental, cujo intuito é demonstrar a validade técnica da solução proposta no Capítulo 3. Com base nos resultados dos experimentos, pretende-se mostrar que a solução proposta encontra o melhor conjunto inicial, é escalável e que a mesma pode ser utilizada em cenários reais, contendo propagações de informações encontradas nos sistemas de redes sociais existentes.

4.1 Objetivos

São vários os objetivos dos experimentos descritos neste capítulo. Em particular, pretende-se avaliar o modelo de Distribuição de Créditos ciente de Tópicos (HCD), em relação às características de acurácia e escalabilidade. Para alcançar esses objetivos, foram conduzidos experimentos para avaliar a acurácia da predição do modelo, a qualidade do conjunto inicial encontrado, o tamanho das propagações obtidas a partir do conjunto inicial e o tempo necessário para encontrar o conjunto inicial.

4.2 Métodos Comparados

A fim de realizar uma análise comparativa dos resultados obtidos com base na utilização do modelo HCD, é necessário definir um conjunto de soluções com as quais os resultados serão comparados. Especificamente, os seguintes modelos foram foco do projeto experimental realizado:

- ***Modelo CD***: primeiro modelo ciente de tópicos proposto neste trabalho, resultante de uma adaptação do modelo de Distribuição de Créditos (CD), originalmente apresentado no trabalho de Goyal *et al.* [35]. No modelo CD, os créditos diretos foram calculados utilizando a Equação (3.2), apresentada no Capítulo 3, desconsiderando os fatores relacionados à Homofilia. Isto é, com o conjunto $H = \emptyset$. Ademais, o

parâmetro λ foi configurado com o valor 0,001 (conforme avaliado e proposto em [35]);

- ***Modelo HCD***: segundo modelo ciente de tópicos proposto neste trabalho, onde os créditos diretos foram calculados utilizando a Equação (3.2), apresentada no Capítulo 3. Além disso, o valor utilizado no parâmetro λ também foi configurado com o valor 0,001 (conforme avaliado e proposto em [35]);

- ***Modelo IC***: implementação do modelo de propagação *Independent Cascade* (IC), no qual as probabilidades das arestas foram aprendidas a partir do conjunto de treinamento. Especificamente, as probabilidades das arestas foram calculadas de acordo com o método *Expected Maximization* (EM), descrito no trabalho de Saito *et al.* [53]. Nos experimentos, foi utilizada a implementação do método EM disponível em Spine [24][1]. Adicionalmente, em todos os experimentos utilizando o modelo IC foram executadas 10.000 simulações MC;

- ***Modelo LT***: implementação do modelo de propagação *Linear Threshold*. Também requer a execução de 10.000 simulações MC. Para aprendizagem das probabilidades das arestas, foram utilizadas algumas ideias apresentadas no trabalho de Kempe *et al.* [34] e Goyal *et al.* [55]. Especificamente, as probabilidades das arestas foram calculadas como: $p_{v,u} = A_{v2u}/N$. Onde A_{v2u} representa o número de ações propagadas do vértice v para o vértice u, dentro do conjunto de treinamento; e N, que corresponde ao número de amigos de v que realizaram a ação antes dele, é utilizado como fator de normalização para assegurar que o somatório das probabilidades sobre as arestas incidentes a v seja 1.

É importante destacar que as versões originais dos modelos IC e LT não são cientes de tópicos. Assim, para comparar esses modelos com a solução proposta neste trabalho, foi utilizada a seguinte metodologia.

1. O conjunto de treinamento, referente ao conjunto de dados contendo propagações relacionadas a um dado Tópico foi utilizado para calcular as probabilidades das arestas dos modelos IC e LT;

2. Foi utilizada uma versão adaptada do algoritmo *Greedy*, com otimização CELF, em conjunto com o respectivo modelo, para encontrar o conjunto inicial.

Adicionalmente aos modelos citados anteriormente, para fins de comparação também foram utilizados alguns algoritmos ou soluções heurísticas consolidados na literatura. Especificamente, foram selecionados os seguintes métodos:

[1]Disponível em: `https://bitbucket.org/mmathioudakis/spine/wiki/Home`

- ***Random***: um método que seleciona de forma aleatória os vértices que irão compor o conjunto inicial;
- ***High Degree (HighDeg)***: um método que seleciona os vértices de maior grau de entrada;
- ***PageRank*** [81]: um método que seleciona os vértices com base na importância do vértice, determinada pelo algoritmo *PageRank*;
- ***HITS*** [82]: um método que seleciona os vértices com base no valor de autoridade calculado para cada vértice.

Para todos os métodos enumerados anteriormente, os k-primeiros vértices foram selecionados para compor o conjunto inicial S, onde k é um parâmetro fornecido como entrada do problema de Maximização de Influência baseado em Tópicos.

4.3 Instrumentação

Nesta seção são apresentadas as bases de dados utilizadas na realização dos experimentos. Além disso, também são descritos os passos realizados para adequar essas bases ao Modelo de Dados descrito no Capítulo 3.

4.3.1 Descrição das Bases de Dados

Para realização dos experimentos deste Capítulo foram escolhidas duas bases de dados disponíveis publicamente na Web: *Epinions* [2] e Flixster[3].

O *Epinions* é um *website* destinado ao público geral que concentra postagens contendo avaliações de produtos feitas por pessoas denominadas de avaliadores. É, portanto, um local onde consumidores podem consultar informações sobre produtos de seu interesse, divididos em 30 categorias (carros, produtos eletrônicos, esportes, etc), e decidir pela compra do produto oferecido a partir do próprio *website*. Além da avaliação do produto em si, informações adicionais sobre os produtos estão disponíveis na forma de comentários realizados por outros usuários que leram a avaliação contida na postagem inicial, ou mesmo, que adquiriram o produto. Após lerem a avaliação disponível e os comentários dos demais usuários, caso os consumidores optem pela compra do produto, eles podem clicar nos *links* contidos na página de avaliação do produto e consultar o preço cobrado pelos diferentes revendedores daquele produto e, eventualmente, concluir o processo de compra.

[2] `http://www.epinions.com/`
[3] `http://www.flixster.com`

Uma característica interessante desse sistema é que, quando os avaliadores publicam postagens sobre os produtos disponíveis, essas avaliações podem ser julgadas como sendo positivas ou negativas pelos demais usuários do sistema. Isso faz com que seja associada, por parte dos consumidores, uma noção de confiança ou desconfiança sobre a avaliação realizada por aquele avaliador. Para explorar essa característica, o *Epinions* disponibiliza uma funcionalidade que possibilita aos consumidores criarem no sistema sua própria rede social, com base na relação de confiança/desconfiança sobre os avaliadores. Dessa forma, os consumidores passam a seguir os avaliadores que eles confiam, e também a receberem diretamente as avaliações produzidas por esses avaliadores.

Por sua vez, o *Flixster* é um sistema de rede social que possibilita aos usuários adicionarem amigos, descobrirem novos filmes e participarem de discussões com outros usuários do sistema sobre filmes ou atores dos quais eles gostam. Antes de começarem a usufruir dessas funcionalidades, os usuários do *Flixster* precisam criar seus próprios perfis no sistema e, em seguida, avaliarem os filmes contidos no catálogo de filmes da rede social e compartilharem essas avaliações com os demais usuários do sistema. Após compartilharem essas avaliações, o sistema começa a sugerir a formação de amizades com outros usuários que possuem gostos similares.

4.3.2 Construção dos Registros de Propagações baseado em Tópicos

Nesta seção são descritos os passos necessários para transformar os conjuntos de dados disponibilizados a partir das bases de dados do *Epinions* e do *Flixster*, respectivamente, no formato definido pela Relação $Actions_T$, que foi definida no Capítulo 3.

Epinions

Neste trabalho, foi utilizada a versão do conjunto de dados do *Epinions*[4] disponibilizada publicamente por Massa & Avessani [97]. Especificamente, esse conjunto de dados encontra-se disponível nos seguintes arquivos:

- No arquivo ***user_rating.txt*** estão armazenados os dados sobre as conexões sociais dos usuários. Especificamente, cada linha deste arquivo está de acordo com o seguinte formato: *user_1*, *user_2*, *value*, *creation*. Uma linha neste formato indica a presença de um relacionamento direcionado (isto é, a presença de uma aresta orientada) entre os usuários *user_1* e *user_2*. O campo *value* (e.g., confiança $= 1$ ou desconfiança $= -1$) indica se o usuário *user_1* confia/desconfia[5] no usuário

[4] `http://www.trustlet.org/wiki/Extended\_Epinions\_dataset`

[5] Apenas os relacionamentos representando a confiança foram selecionados.

user_2; e o campo *creation* indica a data em que esse relacionamento foi criado. As informações deste arquivo são utilizadas para reconstruir um grafo direcionado $RS = (U, R)$, que representa a rede social do *Epinions*;

- No arquivo ***mc.txt*** estão armazenadas algumas informações sobre a autoria dos conteúdos no *Epinions*. Especificamente, em cada linha deste arquivo, estão presentes as informações sobre os identificadores únicos do conteúdo publicado (*content_id*), autor do conteúdo (*author_id*) e assunto (tópico) relacionado ao conteúdo (*subject_id*);

- No arquivo ***rating.txt*** estão armazenadas milhões de entradas de acordo com o seguinte formato: (*content_id*, *member_id*, *rating*, *time*). Especificamente, em cada linha é armazenada uma nota (*rating*) atribuída por um usuário do sistema (*member_id*), ao conteúdo (*content_id*) de autoria de outro usuário, em um dado intervalo de tempo (*time*). A coluna "rating", que armazena as notas atribuídas pelos usuários aos conteúdos publicados, não foi utilizada nos experimentos e, por isso, foi ignorada.

Para que a base do *Epinions* pudesse ser utilizada nos experimentos, foi adotada a seguinte semântica para definir o que seria uma ação realizada pelos usuários no sistema. Em particular, uma ação foi definida como o ato de um usuário avaliar um conteúdo de autoria de outro usuário, que está relacionado a um determinado tópico, em um dado instante de tempo. Dessa forma, quando um usuário u observa outro usuário em que ele confia, por exemplo, v realizar uma ação a, no tempo t_1 e, posteriormente, u também realiza a mesma ação, no tempo t_2, onde $t_1 < t_2$, então, é assumido que u pode ter realizado a ação a sob influência de v. É importante destacar que, se vários usuários em quem u confia, tiverem realizado a mesma ação em um mesmo instante de tempo (anterior à ação realizada pelo usuário u), então não será possível determinar quais desses usuários de fato influenciaram u.

O registro de propagações baseado em tópicos, representado pela relação $Actions_T$(User, Action, Topic, Time), foi construído com base nas informações presentes nos arquivos descritos anteriormente. Desse modo, cada conteúdo distinto representa uma propagação (i.e. uma ação presente na relação $Actions_T$). Consequentemente, o conjunto de usuários que avaliaram um mesmo conteúdo pertencem a uma mesma propagação, onde a ordem de propagação é determinada pelos tempos de realização da ação (ordem cronológica de execução dessas propagações).

Na Tabela 4.1 estão sumarizadas algumas características relacionadas ao conjunto de dados completo do *Epinions*.

Tabela 4.1: Estatísticas do conjunto de dados *Epinions*.

Propriedade	*Epinions*
#Vértices	131580
#Arestas Direcionadas	841372
#Propagações	1560144
#Tuplas	13668319

Flixster

Neste trabalho, foi utilizada a versão do conjunto de dados do *Flixster*[6] disponibilizada publicamente por Jamali & Ester [98]. Especificamente, esse conjunto de dados encontra-se disponível nos seguintes arquivos:

- No arquivo ***links.txt*** estão armazenados os dados sobre as conexões sociais dos usuários. Especificamente, cada linha deste arquivo encontra-se no seguinte formato: $user_1$, $user_2$. Uma linha neste formato indica a existência de um relacionamento direcionado entre dois usuários distintos, representados por $user_1$ e $user_2$, respectivamente. Desse modo, as informações desse arquivo foram utilizadas para reconstruir um grafo direcionado $RS = (U, R)$, que representa a rede social do *Flixster*;

- No arquivo ***users.txt*** estão armazenadas algumas informações sobre os usuários registrados no *Flixster*. Especificamente, em cada linha deste arquivo, estão presentes algumas informações demográficas sobre os usuários, quais sejam: gênero e idade[7]; além do identificador do usuário no sistema. As informações sobre o gênero e a idade são utilizadas no modelo HCD;

- No arquivo ***ratings.txt*** estão armazenadas milhões de entradas, de acordo com o seguinte formato: (*user*, *movie*, *rate*, *time*). Especificamente, em cada linha é armazenada a informação sobre o usuário que avaliou um determinado filme, com uma determinada nota, em um dado intervalo de tempo. A coluna "rate", que é utilizada para armazenar as notas atribuídas pelos usuários aos filmes avaliados (as notas são representadas por valores inteiros que variam no intervalo $[1, 5]$), não foi utilizada nos experimentos e, por isso, foi ignorada;

- No arquivo ***movies.txt*** estão armazenadas as informações sobre o título de cada filme e seu respectivo identificador único no sistema.

De modo similar ao realizado anteriormente na base de dados do *Epinions*, foi adotada a seguinte semântica para definir o que seria uma ação realizada pelos usuários na base

[6] `http://www.cs.ubc.ca/~jamalim/datasets/`

[7] As informações sobre o gênero e a idade são opcionais e não estão disponíveis para todos os usuários

do *Flixster*. Em particular, uma ação foi definida como o ato de um usuário avaliar um filme em um dado instante de tempo. Dessa forma, quando um usuário u observa um de seus amigos, por exemplo, v realizar uma ação a, no tempo t_1 e, posteriormente, u também realiza a mesma ação, no tempo t_2, onde $t_1 < t_2$, então, é assumido que u pode ter realizado a ação a sob influência de v.

O registro de propagações baseado em tópicos, representado pela relação $Actions_T$(User, Action, Topic, Time), foi construído com base nas informações presentes nos arquivos descritos anteriormente. Desse modo, cada filme distinto representa uma ação presente na relação $Actions_T$. Consequentemente, o conjunto de usuários que avaliaram um mesmo filme pertencem a uma mesma propagação, onde a ordem de propagação é determinada pelos tempos de realização da ação.

No entanto, para que a relação $Actions_T$ pudesse ser construída a partir do conjunto de dados do *Flixster*, faltava adicionar a informação sobre o tópico. Como os filmes são relacionados a um ou mais gêneros, as informações sobre os gêneros foram escolhidas para representarem os tópicos. Todavia, tais informações não estão presentes no conjunto de dados original do *Flixster*. Por esse motivo, foi necessária a execução de um processo de *web crawling* sobre alguma base de filmes disponível publicamente na Web, para que essas informações fossem obtidas e, em seguida, adicionadas ao registro de propagações baseado em tópicos. Especificamente, esses dados foram obtidos a partir da base de filmes do *Rotten Tomatoes*, após a execução de requisições à sua API online[8]. Essa base de filmes foi escolhida por ser de propriedade do *Flixster* e, também, por disponibilizar publicamente informações atualizadas sobre seu catálogo de filmes.

É importante destacar que, ao usar a API do *Rotten Tomatoes*, só é possível obter diretamente as informações sobre os gêneros dos filmes, se forem utilizados seus respectivos identificadores únicos. Como os identificadores contidos no conjunto de dados do *Flixster* são diferentes daqueles cadastrados no *Rotten Tomatoes*, uma solução alternativa foi utilizar o título de cada filme como critério de busca para obtenção desses identificadores. Desse modo, para cada filme, foram executadas duas requisições aos serviços da API do *Rotten Tomatoes*. Uma vez de posse das informações sobre os gêneros de cada filme, essas informações foram adicionadas ao arquivo **ratings.txt**. Como resultado final, todas as linhas desse arquivo foram formatadas de modo a corresponderem com a estrutura das tuplas definidas na relação $Actions_T(User, Action, Topic, Time)$.

Na Tabela 4.2 estão sumarizadas algumas características relacionadas ao conjunto de dados completo do *Flixster*.

[8] `http://developer.rottentomatoes.com/`

Tabela 4.2: Estatísticas do conjunto de dados do *Flixster*.

Propriedade	*Flixster*
#Vértices	1002800
#Arestas Direcionadas	11794648
#Propagações	48794
#Tuplas	8196077

4.3.3 Seleção dos Conjuntos de Dados

Conforme descrito anteriormente, um dos objetivos neste capítulo é comparar os modelos HCD, CD, IC e LT utilizando um projeto de experimentos. No entanto, conforme discutido no Capítulo 1, utilizando os modelos de propagação IC e LT, o algoritmo *Greedy* precisa executar simulações Monte Carlo para selecionar o vértice que provê o máximo ganho marginal em cada iteração. Sendo assim, dependendo da quantidade de usuários requerida (parâmetro k) na entrada do problema de Maximização de Influência baseado em Tópicos, podem ser necessários vários dias para encontrar um conjunto inicial S que provê uma solução para o problema.

Nos experimentos conduzidos foram utilizadas apenas as propagações relacionadas aos tópicos com identificadores 526227072 e 462395008 (*Epinions*)[9]; e de Ação e Drama (*Flixster*). Esses tópicos foram escolhidos pelo fato deles serem os mais representativos nos conjuntos de dados do *Epinions* e *Flixster*, respectivamente.

É importante destacar que, para cada tópico selecionado, foi aplicada a seguinte restrição sobre as tuplas da relação $Actions_T(User, Action, Topic, Time)$. Especificamente, optou-se por selecionar apenas os usuários que participaram de pelo menos 20 propagações. Essa restrição foi utilizada para evitar que um usuário, que participou de apenas uma propagação, fosse selecionado para compor o conjunto inicial de algum dos modelos[10].

Adicionalmente, na base do *Flixster*, também optou-se por utilizar apenas as propagações que ocorreram no ano de 2007. Essa restrição se fez necessária, pelo fato de as propagações contidas na base do *Flixster* ocorrerem em períodos de tempo muito diferentes. Por exemplo, algumas propagações começam e terminam dentro de um mesmo ano. Por sua vez, outras propagações podem começar no ano de 2006 e terminar apenas no ano de 2009.

Logo, a restrição de observação a um período de tempo específico, além de permitir a observação do início e término de cada propagação em um mesmo período de tempo, também permite observar como os usuários influenciam os demais participantes da rede social durante esse período. Como consequência, essa decisão contribuiu para uma melhor

[9]No conjunto de dados do *Epinions* não estão disponíveis os significados para os tópicos selecionados. Assim, durante o restante do documento serão utilizados os próprios identificadores para referir-se a cada tópico.

[10]Esse problema foi reportado em [35].

distribuição dos dados, uma redução dos erros nos experimentos relacionados à acurácia dos modelos e a uma melhoria nos resultados desses experimentos.

Em seguida, foram induzidos subgrafos com base nos tópicos selecionados. Esses subgrafos contêm apenas os usuários que participaram de pelo menos 20 propagações relacionadas aos tópicos selecionados. Adicionalmente, também foram separadas do conjunto de dados original do *Epinions* e do *Flixster*, as tuplas do registro de propagações baseado em tópicos que correspondem aos tópicos 526227072 e 462395008; e aos tópicos de Ação e Drama, respectivamente. Como resultado, foram derivados quatro novos conjuntos de dados correspondendo aos referidos tópicos.

Na Tabela 4.3 estão sumarizadas algumas estatísticas para os conjuntos de dados relacionados aos tópicos 526227072 e 462395008, na base de dados do *Epinions*.

Tabela 4.3: Estatísticas dos conjuntos de dados do *Epinions*.

Propriedade	526227072	462395008
#Vértices	1570	1310
#Arestas Direcionadas	111842	118152
Grau Médio	142	180
Diâmetro	6	3
Densidade	0,04540277	0,06890173
#Propagações	8600	3090
#Tuplas	132602	105930

Por sua vez, na Tabela 4.4 estão sumarizadas algumas estatísticas para os conjuntos de dados relacionados aos tópicos de Ação e Drama, na base de dados do *Flixster*.

Tabela 4.4: Estatísticas dos conjuntos de dados do *Flixster*.

Propriedade	Ação	Drama
#Vértices	10074	9542
#Arestas Direcionadas	95992	89672
Grau Médio	19	19
Diâmetro	12	12
Densidade	0,0009459632	0,0009849712
#Propagações	4950	11140
#Tuplas	1181428	1234556

4.4 Experimentos

Nesta seção são descritos os experimentos que foram conduzidos utilizando os conjuntos de dados relacionados aos tópicos das bases de dados do *Epinions* e *Flixster*.

Metodologia

Na realização dos experimentos foi utilizada uma técnica de Validação Cruzada, conhecida por *k-fold Cross Validation* [99]. A técnica em si, consiste em dividir os dados disponíveis

em k-partes distintas, onde cada parte contém aproximadamente a mesma quantidade de dados. Uma vez separadas, cada uma das partes é rotulada com um valor inteiro iniciando de 1 até k, onde k representa a quantidade total de partes (*folds*) disponíveis. A validação dos dados é dita cruzada, pois, em cada validação, uma das partes dos dados é utilizada para compor o **conjunto de teste** e as demais $k-1$ partes são utilizadas para compor o **conjunto de treinamento**. Dessa forma, são executadas k-validações, onde cada parte dos dados é utilizada exatamente uma única vez para compor o conjunto de teste.

Nos experimentos descritos neste capítulo, as propagações relacionadas a um tópico específico foram divididas em cinco partes (i.e. $k = 5$), contendo a mesma quantidade de propagações. Desse modo, para um valor de $k = 5$, foram utilizadas 80% das propagações para compor o conjunto de treinamento e 20% das propagações restantes foram utilizadas para compor o conjunto de teste. Sendo as probabilidades das arestas (modelos IC e LT) e os créditos diretos (modelos CD e HCD) aprendidos com base nas propagações contidas no conjunto de treinamento. Por sua vez, as propagações contidas no conjunto de teste foram utilizadas para avaliar a acurácia dos modelos comparados.

É importante destacar que, ao dividir as propagações em k-partes iguais, tomou-se o cuidado para que uma propagação completa estivesse contida na mesma parte dos dados e que fossem preservadas as distribuições amostrais dos conjuntos de dados originais[11]. Para atender à restrição anterior, foi utilizado um método de amostragem aleatório, sem substituição, a partir do *software* de análise estatística R[12].

Na Tabela 4.5 estão sumarizadas algumas estatísticas sobre os conjuntos de treinamento e de teste relacionados aos tópicos 526227072 e 462395008, na base de dados do *Epinions*.

Tabela 4.5: Conjuntos de dados *Epinions*.

Propriedade	526227072	462395008
#Vértices	1570	1310
#Arestas Direcionadas	111842	118152
#Propagações	8600	3090
#Propagações (Treinamento)	6880	2472
#Propagações (Teste)	1720	618
#Tuplas	132602	105930
#Tuplas (Treinamento)	106188	85548
#Tuplas (Teste)	26414	20382

Por sua vez, na Tabela 4.6 estão sumarizadas algumas estatísticas sobre os conjuntos de treinamento e de teste relacionados aos tópicos de Ação e Drama, na base de dados do *Flixster*.

[11] Nos conjuntos de dados correspondentes aos tópicos de Ação e Drama, por exemplo, foram encontradas distribuições não-normais, com viés positivo, com relação ao tamanho das propagações.

[12] `http://www.r-project.org/`

Tabela 4.6: Conjuntos de dados *Flixster*.

Propriedade	Ação	Drama
#Vértices	10074	9542
#Arestas Direcionadas	95992	89672
#Propagações	4950	11140
#Propagações (Treinamento)	3960	8912
#Propagações (Teste)	990	2228
#Tuplas	1181428	1234556
#Tuplas (Treinamento)	961490	999579
#Tuplas (Teste)	219938	234977

Ambiente de Execução

Todos os experimentos foram executados em uma máquina Intel(R) Core(TM) i7-2670QM CPU @2.20GHz, com 8 GB de memória RAM, Sistema Operacional Linux Mint 17 Cinnamon 64-bits e Kernel do Linux 3.13.0-24-generic. Os algoritmos foram implementados na linguagem de programação C++, com base nos modelos de influência disponibilizados publicamente[13] por Goyal *et al.* [35].

4.4.1 Acurácia dos Modelos

Por meio deste experimento, pretendeu-se compreender quão bons são os métodos comparados em relação à predição do tamanho das propagações observadas no conjunto de teste. Desta forma, neste experimento, foi avaliada a acurácia da predição dos modelos HCD, CD, IC e LT. O resultado deste experimento forneceu evidências iniciais sobre qual é o modelo que possui melhor acurácia. Isto é, cujos resultados estão mais próximos daqueles observados nos conjuntos de dados do *Epinions* e do *Flixster*. Como métrica selecionada para comparação da acurácia dos modelos foi utilizado o *Root Mean Squared Error* (RMSE).

Conforme descrito anteriormente, cada um dos modelos foi submetido a uma etapa de treinamento e outra etapa de teste. Na primeira etapa, as propagações contidas no conjunto de treinamento foram utilizadas para aprendizagem das probabilidades das arestas (no caso dos modelos IC e LT) e dos créditos diretos (no caso dos modelos CD e HCD). Após essa etapa de treinamento, cada um dos modelos foi utilizado para predizer o tamanho das propagações contidas nos conjuntos de testes. A fim de comparar os resultados obtidos neste experimento foi utilizada a seguinte metodologia.

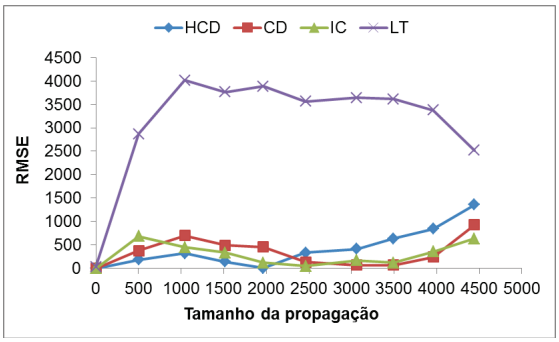
Para um dado conjunto inicial S e um modelo $m \in M = \{CD, HCD, IC, LT\}$, foi computada a propagação esperada, denominada $\sigma_m(S)$. Em seguida, a propagação esperada foi comparada com a propagação real produzida pelo conjunto S. Para calcular a propagação real produzida por um conjunto S, foram realizados os seguintes passos:

[13] `http://www.cs.ubc.ca/~goyal/code-release.php`

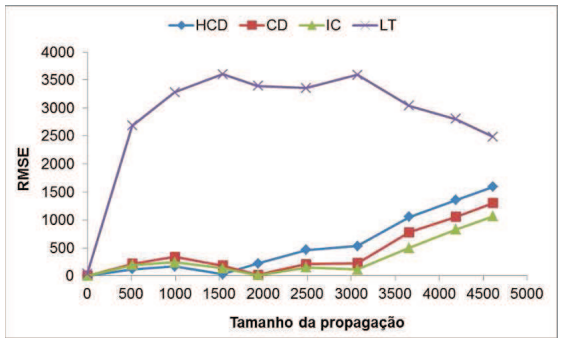
Para cada propagação existente no conjunto de teste, os usuários que foram os primeiros a executarem aquela ação, dentre os seus amigos, foram considerados os iniciadores da propagação. Consequentemente, esses usuários foram escolhidos para compor o conjunto inicial S. Dessa forma, o valor da propagação real (também denominado de tamanho da propagação) produzida pelo conjunto S, pode ser calculado como a quantidade total de usuários que realizaram aquela ação no registro de propagações baseado em tópicos.

Uma vez calculados os valores reais e esperados dos tamanhos das propagações a partir do conjunto de teste, o método utilizado para calcular o erro foi o seguinte: Inicialmente, cada propagação contida no conjunto de teste foi classificada de acordo com o seu valor de $\sigma_m(S)$. O resultado dessa classificação foi a formação de vários grupos[14], onde as propagações contidas em um mesmo grupo possuem o mesmo tamanho. Em seguida, foi computado o RMSE para cada grupo formado anteriormente.

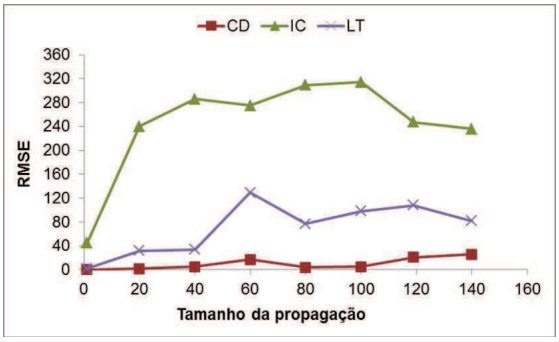
Na Figura 4.1 é apresentado um gráfico de dispersão, no qual pode-se observar o tamanho real das propagações em função do erro (RMSE).



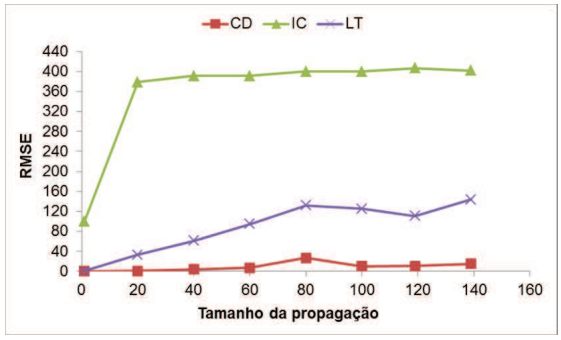
(a) Tópico de Ação (*Flixster*).



(b) Tópico de Drama (*Flixster*).



(c) Tópico 526227072 (*Epinions*).



(d) Tópico 462395008 (*Epinions*).

Figura 4.1: RMSE *versus* Tamanho da propagação.

Com base nos resultados apresentados nas Figuras 4.1(a) e 4.1(b), pode-se observar que os modelos HCD, CD e IC demonstraram ter uma boa acurácia na predição do tamanho das propagações contidas nos respectivos conjuntos de teste, sendo observado os maiores erros para o modelo LT nos experimentos conduzidos com o conjunto de dados do *Flixster*.

[14] A quantidade total de grupos formados depende do conjunto de dados utilizado.

Além disso, pode-se afirmar que, nos intervalos contendo propagações de tamanho 1 a aproximadamente 2000 e propagações de tamanho 1 a aproximadamente 1500, referentes aos tópicos de Ação e Drama, respectivamente; o modelo HCD foi aquele que apresentou os menores valores de RMSE. Ao analisar detalhadamente a quantidade total de propagações contidas nesses intervalos nos dois conjuntos de testes, observa-se que essas propagações correspondem a um total de 96% das propagações relacionadas ao tópico de Ação e 98% das propagações relacionadas ao tópico de Drama.

Esse experimento também foi realizado nos conjuntos de dados referentes aos tópicos 526227072 e 462395008 do *Epinions*. Entretanto, como a base de dados do *Epinions* não disponibiliza quaisquer dados demográficos sobre os usuários (idade e gênero), então o modelo HCD não pôde ser utilizado nos experimentos relacionados aos tópicos desta base de dados.

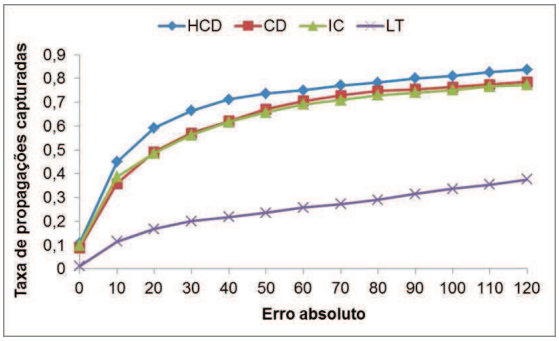
Analisando os resultados apresentados nas Figuras 4.1(c) e 4.1(d), pode-se observar que o modelo CD apresentou os menores valores de RMSE para todos os pontos observados. Esse resultado difere do resultado obtido a partir da base do *Flixster*, uma vez que no *Epinions*, o modelo IC apresentou os maiores erros e o modelo LT foi o segundo modelo de melhor acurácia.

A fim de compreender melhor o resultado anterior, foram analisadas detalhadamente as tuplas contidas nas propagações das bases de dados do *Epinions* e *Flixster*. Com base nos resultados dessa análise, foi observada uma diferença de distribuição dos tempos de realização das ações por parte dos usuários nessas bases de dados. Enquanto no *Epinions* os tempos de realização das ações por parte dos usuários estão muito próximos (uma diferença de poucos minutos), no *Flixster* essa diferença pode ser de vários dias ou até mesmo de meses. Esse resultado é muito interessante, pois exemplifica como o fator de atenuação utilizado na Equação (3.2) impacta diretamente na acurácia do modelo CD e, consequentemente, também na acurácia do modelo HCD.

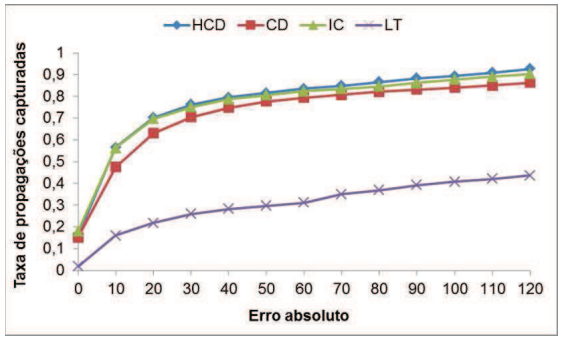
A fim de complementar a análise anterior, foi realizada uma análise adicional, por meio da qual pode-se observar o percentual de propagações capturadas em função do erro absoluto. Os resultados correspondentes a essa análise podem ser visualizados nos gráficos apresentados na Figura 4.2.

Em particular, um ponto (x,y) contendo o valor $(20; 0{,}6)$, representado no gráfico da Figura 4.2(a), é compreendido da seguinte forma: para um valor de erro absoluto de no máximo 20, o modelo HCD consegue capturar 60% das propagações relacionadas ao tópico de Ação contidas no conjunto de teste. Ainda, considerando o mesmo erro absoluto, os modelos CD e IC conseguem capturar metade das propagações, enquanto o modelo LT consegue capturar cerca de 20% das propagações contidas do conjunto de teste.

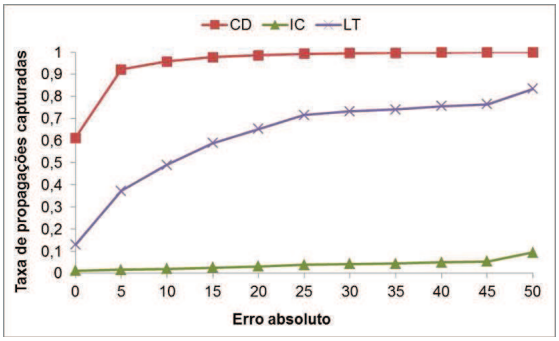
Quando a analise anterior é realizada nos tópicos relacionados à base de dados do



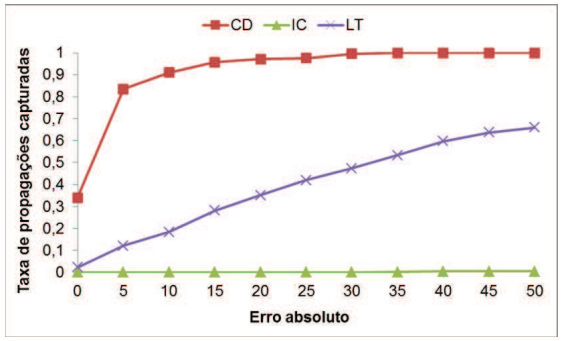
(a) Tópico de Ação (*Flixster*).



(b) Tópico de Drama (*Flixster*).



(c) Tópico 526227072 (*Epinions*).



(d) Tópico 462395008 (*Epinions*).

Figura 4.2: Taxa de propagações capturadas *versus* Erro absoluto.

Epinions, esse resultado torna-se ainda mais evidente em favor do modelo CD, quando o mesmo é comparado com os modelos IC e LT. Por exemplo, considerando um erro absoluto de valor zero, cerca de 60% das propagações são capturadas pelo modelo CD. Adicionalmente, em caso de ser considerado um valor de erro absoluto de apenas 5 (cinco), mais de 90% das propagações são capturadas ao utilizar o modelo CD.

Portanto, com base nos resultados anteriores, pode-se afirmar que os modelos HCD e CD foram os modelos de melhor acurácia nas bases de dados do *Flixster* e *Epinions*, respectivamente, pois foram aqueles que apresentaram os menores erros nas suas predições no intervalo contendo a maior parte dos dados dos conjuntos de testes.

Por fim, também é importante destacar que pelo fato de o modelo HCD ser uma extensão do modelo de Distribuição de Créditos (CD), era de se esperar que esses modelos tivessem acurácia muito próxima. Entretanto, especialmente no conjunto de dados contendo propagações relacionadas ao tópico de Ação, a diferença de acurácia entre os modelos foi mais evidenciada em favor do modelo HCD. Esse resultado fornece evidências iniciais de que a adição das informações relacionadas à Homofilia na Equação de distribuição de créditos diretos produziu uma melhoria na acurácia das predições dos modelos ciente de tópicos baseados em distribuição de créditos.

4.4.2 Similaridade entre os Conjuntos Iniciais

No segundo experimento foi avaliada a similaridade entre os conjuntos iniciais encontrados pelo algoritmo *Greedy*, utilizando os modelos HCD, CD, IC e LT, respectivamente. Além desses modelos, também foram adicionados ao experimento conduzido os seguintes algoritmos e heurísticas: *Random*, *High Degree*, *PageRank* e *HITS*.

Utilizando cada um dos modelos enumerados anteriormente, foram selecionados até 50 usuários ($k = 50$) para compor o conjunto inicial S. O conjunto selecionado contém apenas usuários que realizaram propagações relacionadas aos tópicos 526227072 e 462395008 (*Epinions*); e de Ação e Drama (*Flixster*).

A fim de calcular a similaridade entre os conjuntos iniciais encontrados a partir da utilização de dois modelos foi utilizado o método descrito a seguir:

Sejam $X = \{x_1,x_2,\ldots,x_n\}$ e $Y = \{y_1,y_2,\ldots,y_n\}$ dois vetores de tamanho N, indexados de $1 \ldots N$, representando os conjuntos iniciais encontrados por dois métodos distintos, por exemplo, HCD e CD. Então, a similaridade entre os dois vetores pode ser calculada utilizando a Equação (4.1):

$$sim(X,Y) = \begin{cases} 0 & \text{se } |W| = 0; \\ \left(1 - \left(\frac{1}{N(N-1)} \sum_{i=1}^{N} |a_i - b_i|\right)\right) \frac{|W|}{N} & \text{se } |W| > 0. \end{cases} \tag{4.1}$$

Onde:

- O vetor $W = \{w_1,w_2,\ldots,w_k\}$ $|w_i \in X \wedge w_i \in Y$, $k \leq N$, armazena os elementos comuns aos vetores X e Y;
- Os vetores $A = \{a_1,a_2,\ldots,a_n\}$ e $B = \{b_1,b_2,\ldots,b_n\}$ armazenam os *rankings* (posições) dos elementos contidos no vetor W, referentes aos vetores X e Y, respectivamente. De forma que, $\forall i$ $|a_i - b_i| \leq N - 1$ e $\sum_{i=1}^{N} |a_i - b_i| \leq N(N-1)$. Por fim, o *ranking* dos elementos que não fazem parte de W, isto é, $(X \cup Y) - W$, é igual a zero.

A função definida anteriormente utiliza como base duas propriedades importantes para o cálculo da similaridade entre dois vetores de mesmo tamanho. A primeira confere importância à posição de cada elemento nos vetores originais. Para isso, no cálculo de similaridade são penalizadas as diferenças entre essas posições para os elementos contidos em W, com base na Distância de *Manhatthan* [100]. Por sua vez, a segunda propriedade confere importância à quantidade total de interseções entre os elementos dos dois vetores.

Desse modo, considerando as propriedades anteriores, a utilização da função de similaridade produzirá como resultado um valor real definido no intervalo $[0, 1]$, onde o valor

zero indica que os elementos presentes nos vetores são completamente distintos. E, o valor 1 indica uma correspondência perfeita entre esses vetores, em relação à quantidade de interseções e respectivas posições de seus elementos.

Nas Tabelas 4.7 e 4.8 estão sumarizados o número de interseções encontradas e o valor da similaridade (valor dentro dos parênteses) entre os conjuntos iniciais dos modelos comparados, para as propagações relacionadas aos tópicos de Ação e Drama, respectivamente.

Tabela 4.7: Análise de similaridade entre os conjuntos iniciais para propagações relacionadas ao tópico Ação.

Modelo	HCD	CD	IC	LT	*Random*	*HighDeg*	*PageRank*	*HITS*
HCD	50 (1,0)	40 (0,69)	2 (0,04)	35 (0,57)	0 (0)	9 (0,15)	12 (0,21)	2 (0,04)
CD	-	50 (1,0)	2 (0,04)	33 (0,57)	0 (0)	6 (0,11)	9 (0,17)	2 (0,04)
IC	-	-	50 (1,0)	4 (0,08)	0 (0)	5 (0,1)	7 (0,13)	2 (0,04)
LT	-	-	-	50 (1,0)	0 (0)	8 (0,15)	11 (0,2)	2 (0,04)
Random	-	-	-	-	50 (1,0)	0 (0)	0 (0) (0)	0 (0)
HighDeg	-	-	-	-	-	50 (1,0)	39 (0,67)	15 (0,27)
PageRank	-	-	-	-	-	-	50 (1,0)	6 (0,11)
HITS	-	-	-	-	-	-	-	50 (1,0)

Tabela 4.8: Análise de similaridade entre os conjuntos iniciais para propagações relacionadas ao tópico Drama.

Modelo	HCD	CD	IC	LT	*Random*	*HighDeg*	*PageRank*	*HITS*
HCD	50 (1,0)	39 (0,7)	3 (0,06)	31 (0,52)	0 (0)	8 (0,15)	10 (0,18)	2 (0,04)
CD	-	50 (1,0)	2 (0,04)	34 (0,57)	0 (0)	8 (0,15)	9 (0,17)	1 (0,02)
IC	-	-	50 (1,0)	4 (0,08)	0 (0)	3 (0,06)	4 (0,08)	3 (0,06)
LT	-	-	-	50 (1,0)	0 (0)	9 (0,17)	11 (0,21)	2 (0,04)
Random	-	-	-	-	50 (1,0)	0 (0)	0 (0)	0 (0)
HighDeg	-	-	-	-	-	50 (1,0)	42 (0,74)	4 (0,08)
PageRank	-	-	-	-	-	-	50 (1,0)	3 (0,06)
HITS	-	-	-	-	-	-	-	50 (1,0)

Com base nos resultados anteriores, pode-se observar um valor de similaridade bastante significativo (a seleção de 31 a 40 usuários iguais e um valor de similaridade de pelo menos 0,52) entre os conjuntos iniciais dos modelos HCD, CD e LT, nas propagações relacionadas aos tópicos de Ação e Drama. Por sua vez, a similaridade entre os conjuntos iniciais dos modelos HCD, CD e LT com o conjunto inicial do modelo IC foi muito pequena (entre 2 a 4 usuários e um valor de similaridade de no máximo 0,08).

Ademais, os conjuntos iniciais dos modelos HCD, CD e LT também apresentaram um valor de similaridade significativo com os conjuntos iniciais selecionados utilizando a heurística *HighDeg* (pelo menos 0,15) e o método *PageRank* (pelo menos 0,17), nas propagações relacionadas aos tópicos de Ação e Drama. Esse resultado é interessante, pois demonstra que usuários importantes, com grande quantidade de conexões, estão ligados entre si. Esse último fato favorece à propagação de informações para os demais usuários

dessas redes sociais. Todavia, outros fatores ainda não observados parecem contribuir para a escolha dos usuários que farão parte do conjunto inicial.

Adicionalmente, também pode-se observar um valor de similaridade de pelo menos 0,67 entre os conjuntos iniciais selecionados pela heurística *HighDeg* e pelo método *PageRank*. A similaridade entre os conjuntos iniciais dos demais modelos não foi significativa.

Por fim, apesar de o modelo HCD ser baseado no modelo CD, a equação de distribuição de créditos diretos utilizando Homofilia, utilizada pelo modelo HCD, produziu um impacto na forma como o algoritmo *Greedy* selecionou os usuários pertencentes aos conjuntos iniciais desses dois modelos, nos tópicos de Ação e Drama (Tabelas 4.7 e 4.8). Especificamente, foi encontrada uma diferença de cerca de 20% (cerca de 10 usuários diferentes dentre os 50 usuários selecionados) entre os elementos dos conjuntos iniciais dos modelos HCD e CD. O impacto dessa diferença sobre o tamanho da propagação produzida pelos respectivos conjuntos iniciais foi avaliada no terceiro experimento conduzido, descrito na seção 4.4.3.

Por sua vez, nas Tabelas 4.9 e 4.10 estão sumarizados o número de interseções encontradas e o valor da similaridade entre os conjuntos iniciais dos modelos comparados, para as propagações relacionadas aos tópicos 526227072 e 462395008, respectivamente.

Tabela 4.9: Análise de similaridade entre os conjuntos iniciais para propagações relacionadas ao tópico 526227072.

Modelo	CD	IC	LT	*Random*	*HighDeg*	*PageRank*	*HITS*
CD	50 (1,0)	1 (0,02)	18 (0,33)	3 (0,06)	14 (0,25)	13 (0,23)	17 (0,29)
IC	-	50 (1,0)	0 (0)	1 (0,02)	2 (0,04)	2 (0,04)	2 (0,04)
LT	-	-	50 (1,0)	0 (0)	18 (0,32)	17 (0,3)	18 (0,32)
Random	-	-	-	50 (1,0)	1 (0,02)	1 (0,02)	2 (0,04)
HighDeg	-	-	-	-	50 (1,0)	48 (0,9)	37 (0,66)
PageRank	-	-	-	-	-	50 (1,0)	35 (0,61)
HITS	-	-	-	-	-	-	50 (1,0)

Tabela 4.10: Análise de similaridade entre os conjuntos iniciais para propagações relacionadas ao tópico 462395008.

Modelo	CD	IC	LT	*Random*	*HighDeg*	*PageRank*	*HITS*
CD	50 (1,0)	0 (0)	23 (0,4)	1 (0,002)	18 (0,3)	16 (0,28)	20 (0,35)
IC	-	50 (1,0)	0 (0)	2 (0,04)	4 (0,08)	4 (0,08)	3 (0,06)
LT	-	-	50 (1,0)	1 (0,02)	20 (0,35)	19 (0,34)	20 (0,35)
Random	-	-	-	50 (1,0)	1 (0,02)	1 (0,02)	1 (0,02)
HighDeg	-	-	-	-	50 (1,0)	47 (0,9)	39 (0,7)
PageRank	-	-	-	-	-	50 (1,0)	36 (0,64)
HITS	-	-	-	-	-	-	50 (1,0)

Com base nos resultados anteriores, os modelos CD e LT apresentaram um valor de similaridade menos significativo do que aquele encontrado na base do *Flixster* (a seleção de 18 a 23 usuários iguais e um valor de similaridade de pelo menos 0,33) entre os conjuntos

iniciais dos modelos CD e LT, nas propagações relacionadas aos tópicos 526227072 e 462395008. Por sua vez, a similaridade entre os conjuntos iniciais dos modelos CD e LT com o conjunto inicial do modelo IC foi praticamente inexistente (apenas 0,02).

Todavia, dois resultados chamaram muito a atenção. O primeiro deles foi o valor de similaridade dos modelos CD e LT com os métodos *HighDeg* (mínimo de 0,25), *PageRank* (mínimo de 0,23) e *HITS* (mínimo de 0,29) nas propagações relacionadas aos tópicos 526227072 e 462395008. O segundo foi o alto valor de similaridade entre os métodos *HighDeg*, *PageRank* e *HITS*.

Adicionalmente, também foi avaliada a similaridade entre os conjuntos iniciais encontrados pelo algoritmo *Greedy*, fixando-se um mesmo modelo (por exemplo, HCD com HCD ou CD com CD), para os conjuntos de dados contendo propagações relacionadas aos tópicos 526227072 e 462395008; e Ação e Drama. Nas Tabelas 4.11 e 4.12 são apresentados os resultados dessa análise.

Tabela 4.11: Análise de similaridade dos conjuntos iniciais para propagações relacionadas aos tópicos 526227072 e 462395008.

526227072	462395008	Similaridade
CD	CD	28 (0,49)
IC	IC	2 (0,04)
LT	LT	22 (0,41)
Random	*Random*	0 (0)
HighDeg	*HighDeg*	39 (0,7)
PageRank	*PageRank*	38 (0,68)
HITS	*HITS*	43 (0,79)

Tabela 4.12: Análise de similaridade dos conjuntos iniciais para propagações relacionadas aos tópicos de Ação e Drama.

Ação	Drama	Similaridade
HCD	HCD	42 (0,75)
CD	CD	42 (0,74)
IC	IC	27 (0,49)
LT	LT	36 (0,65)
Random	*Random*	0 (0)
HighDeg	*HighDeg*	38 (0,69)
PageRank	*PageRank*	43 (0,79)
HITS	*HITS*	39 (0,72)

Com base nos resultados apresentados nas Tabelas 4.11 e 4.12, observa-se que em todos os conjuntos iniciais houve uma interseção de pelo menos 60% entre os usuários para os tópicos avaliados. As únicas exceções foram o método *Random*, que selecionou conjuntos iniciais distintos para os tópicos considerados; e o modelo IC, que na base de dados do *Epinions* apresentou interseção muito baixa dos conjuntos iniciais nos tópicos avaliados. Adicionalmente, esses resultados fornecem evidências de que:

- A influência social exercida pelos usuários depende dos tópicos avaliados. Desse modo, mesmo que os usuários possam exercer influência sobre seus amigos em mais de um tópico, essa é a exceção e não a regra;

- Os usuários selecionados possuem influência social distinta em cada tópico. Por exemplo, conforme apresentado na Tabela 4.13, apenas três usuários dentre os dez mais influentes no tópico 526227072 também fazem parte da lista dos dez usuários mais influentes no tópico 462395008.

Tabela 4.13: Lista dos 10 usuários mais influentes nos tópicos 526227072 e 462395008, utilizando o modelo CD.

Usuário	*Ranking* (526227072)	*Ranking* (462395008)
3270341	1	-
3089094	2	-
3188374	3	-
3236151	**4**	**8**
3018333	5	24
3138171	**6**	**9**
3385398	**7**	**6**
3234187	8	23
3346032	9	37
3341035	10	35

- O resultado anterior foi menos evidente no *Flixster*, conforme sumarizado na Tabela 4.14, onde houve uma correspondência maior entre os *rankings* dos usuários mais influentes nos tópicos de Ação e Drama. Todavia, ainda assim, podem ser observadas algumas diferenças significativas nos *rankings* dos usuários 16469, 696615 e 567205.

Tabela 4.14: Lista dos 10 usuários mais influentes nos tópicos de Ação e Drama, utilizando o modelo HCD.

Usuário	*Ranking* (Ação)	*Ranking* (Drama)
35778	1	2
917304	2	3
923538	3	4
16469	**4**	**11**
58055	5	8
911133	6	6
696615	**7**	**1**
594306	8	9
567205	**9**	**5**
12660	10	10

Os resultados anteriores corroboram com a teoria do comportamento social coletivo de Granovetter [57].

4.4.3 Tamanho da Propagação produzida pelo Conjunto Inicial

No terceiro experimento conduzido, o objetivo foi avaliar dentre as soluções encontradas pelos métodos HCD, CD, IC e LT, *HighDeg*, *PageRank* e *HITS*, quais delas produziriam o maior tamanho de propagações relacionadas aos tópicos 526227072 e 462395008 (*Epinions*); e aos tópicos de Ação e Drama (*Flixster*).

A fim de comparar o tamanho das propagações produzidas pelos conjuntos iniciais de cada modelo, foi necessário primeiramente determinar um modelo comum, denominado **modelo de referência**, onde a propagação produzida por cada conjunto inicial pudesse ser comparada. Isso se fez necessário, pois para um conjunto inicial de tamanho arbitrário não é conhecido *a priori* os valores reais de propagação produzidos por esse conjunto. Por esse motivo, uma segunda solução possível seria eleger um dos modelos avaliados como o modelo de referência e, utilizando este modelo, comparar as saídas produzidas pelos conjuntos iniciais dos demais modelos. Uma vez que, os modelos HCD e CD foram aqueles que apresentaram melhor acurácia, nas bases de dados do *Flixster* e *Epinions*, respectivamente, eles foram escolhidos como os modelos de referência para execução dos experimentos nessas bases.

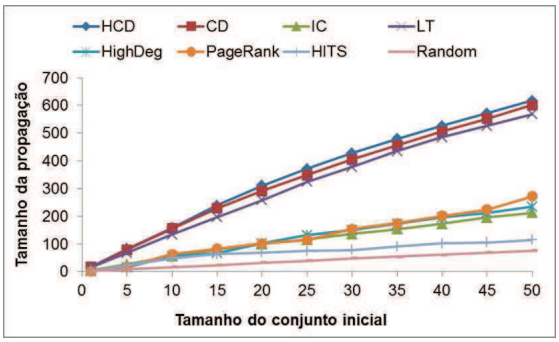
Assim, a metodologia utilizada neste experimento foi a seguinte: Para cada modelo comparado, uma instância do conjunto inicial S, onde $|S| = k$, foi fornecida isoladamente como entrada para o modelo de referência. Sendo, em seguida, observado o tamanho da propagação produzida por esse conjunto inicial. Esse procedimento foi repetido, em cada base de dados, variando-se o valor do parâmetro k, em cinco unidades inteiras, no intervalo $[1, 50]$. Os resultados são apresentados na Figura 4.3.

Conforme pode ser observado nas Figuras 4.3(a) e 4.3(b), referentes aos conjuntos de dados do *Flixster*, contendo apenas propagações relacionados tópicos de Ação e Drama, respectivamente, os conjuntos iniciais de usuários encontrados com base nos modelos HCD, CD e LT foram os que propagaram informações para a maior quantidade de usuários.

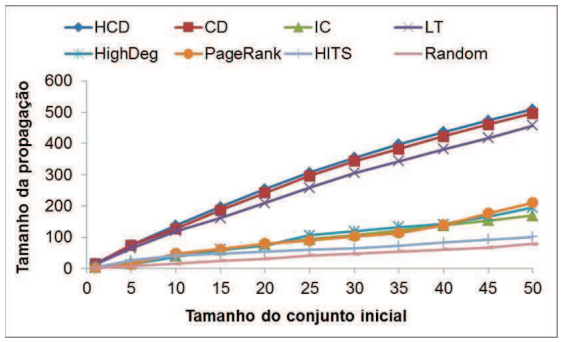
Por sua vez, nas Figuras 4.3(c) e 4.3(d), referentes aos conjuntos de dados do *Epinions*, contendo propagações relacionadas aos tópicos 526227072 e 462395008, respectivamente, os conjuntos iniciais de usuários encontrados com base nos modelos CD e LT foram os que propagaram informações para a maior quantidade de usuários.

Estes resultados são consistentes com o experimento 2, descrito na seção 4.4.2, onde foi observada uma similaridade significativa entre os conjuntos iniciais dos modelos HCD, CD e LT no *Flixster*; e entre os conjuntos iniciais dos modelos CD e LT no *Epinions*.

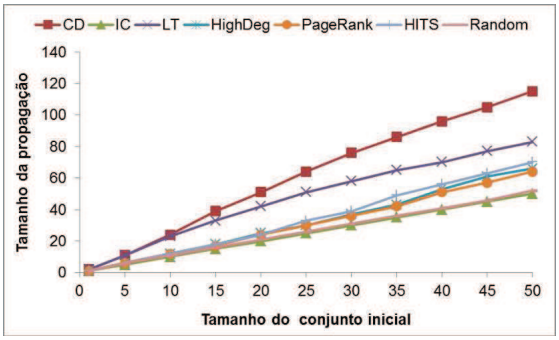
No experimento 2, descrito na seção 4.4.2, foi encontrada no *Flixster* uma diferença de cerca de 20% entre os usuários contidos nos conjuntos iniciais dos modelos HCD e CD. Analisando esses dados detalhadamente, em conjunto com os gráficos apresentados na Figura 4.3, observa-se que essa diferença está localizada próxima às regiões onde o



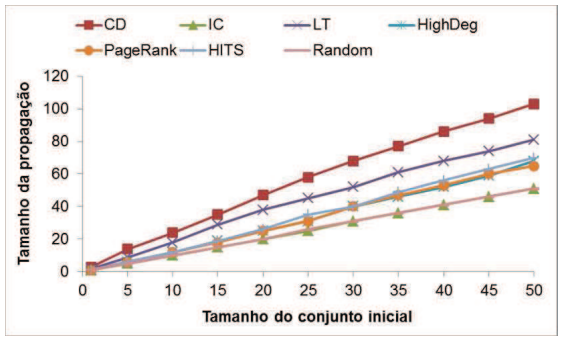
(a) Tópico de Ação (*Flixster*).



(b) Tópico de Drama (*Flixster*).



(c) Tópico 526227072 (*Epinions*).



(d) Tópico 462395008 (*Epinions*).

Figura 4.3: Comparação do tamanho das propagações produzidas pelos conjuntos iniciais de usuários.

parâmetro $k \geq 15$ (Ação) e $k \geq 10$ (Drama), respectivamente. Avaliando o impacto dessa diferença sobre o tamanho total da propagação, observa-se que houve um aumento de 3% em favor do modelo HCD em relação ao modelo CD.

A princípio um aumento de 3% sobre o tamanho total de uma propagação não parece ser tão substancial. Entretanto, é importante salientar que as redes sociais possuem milhões ou mesmo bilhões de usuários registrados e esse percentual representaria a adição de potenciais interessados em um determinado tipo de informação/produto/serviço. Assim, considerando um sistema de rede social com algumas dezenas de milhões de usuários, se um anunciante da empresa fictícia ACME atualmente consegue propagar um anúncio sobre um novo *smartphone* para 10 milhões de pessoas potencialmente interessadas nesse tipo de produto. Então, um aumento de 3% sobre o tamanho total dessa propagação poderá resultar na adição de 300 mil novos potenciais consumidores desse tipo produto.

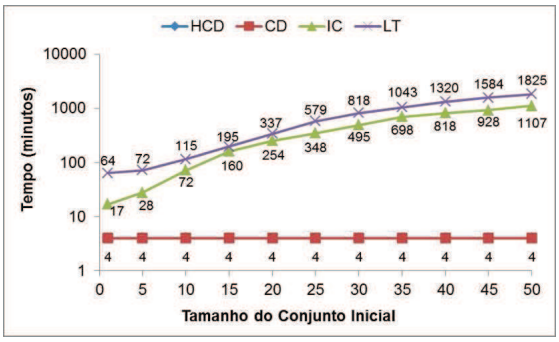
4.4.4 Tempo de Execução

No quarto experimento conduzido, o objetivo foi avaliar o tempo necessário para que uma instância do algoritmo *Greedy* (com implementação CELF), utilizando como base os modelos HCD, CD, IC e LT, encontrasse um conjunto inicial S com tamanho k.

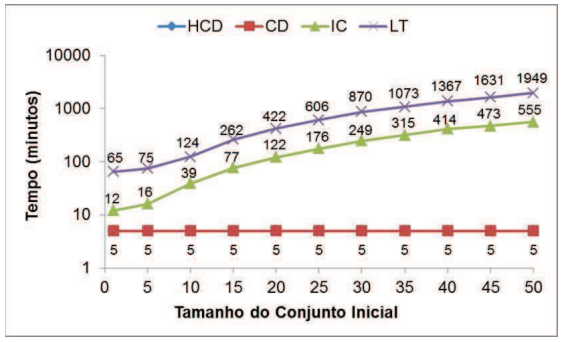
Assim, para cada modelo sendo comparado, uma instância do algoritmo *Greedy* foi

executada sobre um grafo representando as conexões sociais presentes nos conjuntos de dados contendo propagações relacionadas aos tópicos de Ação e Drama (*Flixster*); e aos tópicos 526227072 e 462395008 (*Epinions*), respectivamente. Adicionalmente, também foi fornecido à essa instância do algoritmo *Greedy* um parâmetro k, que representa a quantidade de usuários presentes no conjunto inicial. Como resultado, foi encontrado um conjunto inicial S, onde $|S| = k$. Cada experimento foi repetido, variando-se o valor do parâmetro k, em cinco unidades inteiras, dentro do intervalo $[1, 50]$.

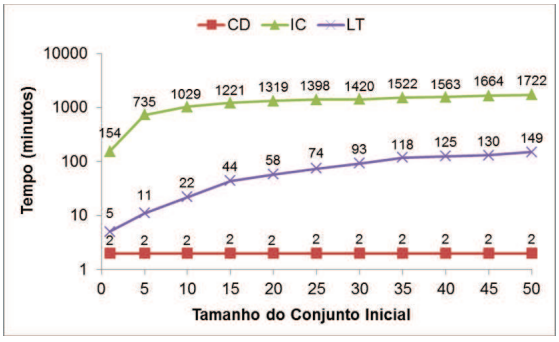
Os resultados dos experimentos são apresentados na Figura 4.4, em escala logarítmica, para possibilitar uma melhor observação das diferenças de tempo entre os modelos comparados. Nas Figuras 4.4(a), 4.4(b), 4.4(c) e 4.4(d) são apresentados os resultados para os conjuntos de dados correspondentes às propagações relacionadas aos tópicos de Ação e Drama (*Flixster*)[15]; e aos tópicos 526227072 e 462395008 (*Epinions*), respectivamente.



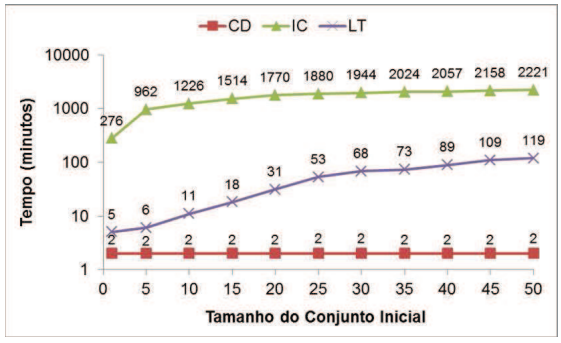
(a) Tópico de Ação (*Flixster*).



(b) Tópico de Drama (*Flixster*).



(c) Tópico 526227072 (*Epinions*).



(d) Tópico 462395008 (*Epinions*).

Figura 4.4: Comparativo do tempo de execução para os modelos HCD, CD, IC e LT.

Analisando os gráficos apresentados na Figura 4.4, observa-se que, à medida que o parâmetro k aumenta, o algoritmo *Greedy*, utilizando como base os modelos IC ou LT, torna-se muito ineficiente para encontrar um conjunto inicial com tamanho k. Considere, como exemplo, as propagações relacionadas ao tópico de Ação na base de dados do *Flixster*. Para esse conjunto de dados, o algoritmo *Greedy*, utilizando como base o modelo de

[15]Nas Figuras 4.4 (a) e 4.4 (b), os resultados do tempo de execução para os modelos HCD e CD foram os mesmos. Logo, as séries estão sobrepostas.

propagação IC, necessitou de cerca de 18 horas e 30 minutos para encontrar um conjunto inicial contendo 50 usuários. Considerando ainda as propagações relacionadas à esse mesmo tópico, o tempo de execução do algoritmo *Greedy*, utilizando o modelo LT, foi ainda pior, uma vez que foi necessário esperar 1 dia e 6 horas para encontrar um conjunto inicial contendo 50 usuários. De forma contrária, o algoritmo *Greedy*, utilizando como base os modelos CD ou HCD, necessitou de apenas 4 minutos para completar a mesma tarefa.

Quando são comparados os resultados referentes aos conjuntos de dados do *Epinions* com aqueles observados na base de dados do *Flixster*, observa-se que o tempo de execução do algoritmo *Greedy*, utilizando como base o modelo LT, foi muito menor do que o tempo de execução de outra instância do algoritmo *Greedy*, utilizando como base o modelo IC. Conforme discutido anteriormente, no experimento referente à acurácia dos modelos, esse fato se deve há uma diferença de distribuição dos tempos de realização das ações por parte dos usuários nessas bases de dados. Assim, da mesma forma que essa diferença ocasionou naquele experimento uma piora sensível na acurácia do modelo IC, ela também fez com que, neste experimento, o algoritmo *Greedy* demorasse mais tempo para selecionar o usuário que deveria ser adicionado ao conjunto inicial S, em cada uma das 50 iterações.

Com base nos resultados anteriores, pode-se afirmar que, tanto no *Flixster* quanto no *Epinions*, o algoritmo *Greedy*, utilizando como base os modelos CD ou HCD, consegue determinar rapidamente uma solução para o problema de Maximização de Influência baseado em Tópicos. E, que esses métodos são várias ordens de magnitude mais rápidos do que os modelos IC e LT. Esses resultados já eram esperados, uma vez que os modelos CD e HCD não requerem a realização de uma etapa para aprender as probabilidades das arestas e, tampouco, necessitam executar simulações Monte Carlo para selecionar os usuários do conjunto inicial S, como ocorre nos modelos IC e LT.

4.5 Considerações Finais

Neste capítulo foi realizada uma análise experimental para avaliar o impacto da utilização dos modelos de Distribuição de Créditos ciente de Tópicos e do registro de propagações baseado em tópicos no problema de Maximização de Influência baseado em Tópicos.

Especificamente, utilizando conjuntos de dados contendo informações sobre a estrutura social e histórico de propagações de dois sistemas de redes sociais reais, foi conduzido um conjunto de experimentos para avaliar os modelos HCD, CD, IC e LT, em relação às seguintes características: (i) acurácia do modelo, (ii) similaridade dos conjuntos iniciais encontrados, (iii) tamanho das propagações produzidas por cada conjunto inicial e (iv) o tempo necessário para o algoritmo *Greedy* encontrar o conjunto inicial utilizando cada

modelo.

Com base nos principais resultados apresentados, foram fornecidas evidências de que: (1) os modelos de distribuição de créditos ciente de tópicos possuem melhor acurácia do que os modelos de propagação IC e LT (HCD na base do *Flixster* e CD na base do *Epinions*); (2) os conjuntos iniciais encontrados pelo algoritmo *Greedy*, utilizando os modelos comparados neste capítulo, são diferentes; (3) os conjuntos iniciais encontrados utilizando os modelos de distribuição de créditos ciente de tópicos produzem as maiores propagações (HCD na base do *Flixster* e CD na base do *Epinions*); e (4) o modelo HCD é equiparável ao modelo CD, em relação ao tempo necessário para encontrar os k-usuários do conjunto inicial. Além disso, os dois modelos são ordens de magnitude mais rápidos do que os modelos que utilizam simulações Monte Carlo.

Como resultado adicional, também foram encontradas evidências de que a influência social exercida pelos usuários é dependente do tópico que está sendo considerado, e também que os usuários exercem influência social distinta em tópicos diferentes. Esses resultados reforçam algumas das suposições encontradas na teoria sobre o comportamento social coletivo, presentes no trabalho de Granovetter [57].

Capítulo 5

Conclusões

Nesta tese abordou-se o problema de Maximização de Influência baseado em Tópicos. Esse problema de otimização discreta consiste em selecionar um conjunto de usuários que sejam capazes de propagar informações relacionadas a um tópico de interesse para uma parcela substancial de usuários de uma rede social.

Do ponto de vista técnico, mudanças constantes na estrutura das redes sociais (por exemplo, entrada/saída de usuários, surgimento/término de relacionamentos sociais), em conjunto com a escala dos sistemas de redes sociais (medida pela quantidade total de usuários e de conexões sociais registradas no sistema) contribuem fortemente para o aumento da complexidade de encontrar de forma rápida uma solução eficiente para o problema de Maximização de Influência baseado em Tópicos. Ao considerar tais desafios, é necessário que soluções para esse problema reflitam rapidamente as mudanças na estrutura da rede social, dos tópicos de interesse dos usuários, e também permitam que a influência social, exercida pelos usuários em cada tópico, seja (re)aprendida rapidamente, enquanto é mantida a qualidade do conjunto inicial de usuários encontrado.

Ao analisar as soluções existentes, observou-se que tais soluções não são adequadas para redes sociais de larga escala e precisam incorporar mecanismos para determinar a influência social exercida entre os usuários em relação a cada tópico de interesse. Entretanto, conforme discutido neste documento, a utilização desses mecanismos torna essas soluções não escaláveis. Consequentemente, para estas abordagens, torna-se difícil ou mesmo inviável lidar de forma rápida e eficiente com as mudanças constantes na estrutura das redes sociais. Por outro lado, as soluções que (re)aprendem as informações sobre a influência social a partir do histórico de propagações e que não necessitam de simulações Monte Carlo para selecionar os usuários que farão parte do conjunto inicial, são capazes de lidar com a escala, a dinamicidade inerente aos sistemas de redes sociais e ainda conseguem encontrar conjuntos iniciais eficientes. Todavia, nenhuma dessas soluções era ciente dos tópicos de interesse dos usuários.

Motivado por este problema, no contexto desta tese, foi concebida uma solução escalável para maximizar a propagação de informações em redes sociais *online*, com base na influência social, e ciente dos tópicos de interesse dos usuários. Mais especificamente, a solução apresentada, a partir de um grafo representando uma rede social, um histórico de propagações de informações, um tópico de interesse e um parâmetro k representando a quantidade de usuários a serem encontrados, permite: (i) inferir dinamicamente os tópicos de interesse dos usuários a partir de dados de propagações reais; (ii) inferir o nível de influência social entre os usuários, considerando tópicos de interesses similares; e (iii) minerar diretamente um conjunto de k-usuários que maximiza a propagação de informações na rede social por tópico.

5.1 Contribuições

A seguir apresenta-se um resumo das principais contribuições do trabalho no contexto do problema de Maximização de Influência baseado em Tópicos:

- Definição de um modelo para representar o histórico de propagações realizadas pelos usuários em uma rede social;
- Definição de um modelo para aprendizagem de influência social por tópicos de interesse dos usuários, a partir de registros de propagações reais;
- Especificação e desenvolvimento de uma solução escalável, com garantia de aproximação em relação à solução ótima, para encontrar os k-usuários que maximizam a propagação de informações em redes sociais, de acordo com a influência social exercida pelos usuários em cada tópico de interesse e as características dos conteúdos propagados na rede social;
- Realização de uma validação técnica da solução. Especificamente, avaliou-se a acurácia da predição do modelo, a qualidade do conjunto inicial encontrado, o tamanho das propagações obtidas a partir do conjunto inicial e o tempo necessário para encontrar o conjunto inicial, em relação às soluções encontradas no estado da arte.

Em particular, as três primeiras contribuições encontram-se no Capítulo 3. Especificamente, na Seção 3.4, foi apresentado o Modelo de Distribuição de Créditos ciente de Tópicos (modelo HCD). O modelo HCD permite aprender a influência de cada usuário de acordo com um tópico específico. Neste modelo, também foi introduzido o conceito de registro de propagações baseado em tópicos e foram definidas equações para calcular os créditos diretos e totais dados a um conjunto de usuários, devido aos mesmos terem

influenciado seus amigos em várias ações relacionadas a um dado tópico. Além disso, na equação utilizada para distribuir os créditos diretos, foram adicionadas informações relacionadas a homofilia.

Em seguida, nas Seções 3.5 e 3.6 foi apresentada uma solução de mineração direta de dados, que utiliza as propagações existentes no registro de propagações baseado em tópicos para alimentar o modelo HCD. Tal solução não necessita aprender as probabilidades das arestas e, tampouco, em cada iteração do algoritmo *Greedy*, na etapa de seleção dos usuários, executar simulações MC para calcular o ganho marginal provido por um usuário em relação ao conjunto inicial. Por fim, foram descritos os algoritmos utilizados para processar de forma eficiente o registro de propagações baseado em tópicos e encontrar rapidamente o conjunto inicial procurado no problema abordado nesta tese.

Por sua vez, a última contribuição encontra-se no Capítulo 4. Com base nos principais resultados apresentados, foram fornecidas evidências de que: (1) os modelos de distribuição de créditos ciente de tópicos possuem melhor acurácia do que os modelos de propagação IC e LT (HCD na base do *Flixster* e CD na base do *Epinions*); (2) os conjuntos iniciais encontrados pelo algoritmo *Greedy*, utilizando os modelos CD, HCD, IC e LT são diferentes; (3) os conjuntos iniciais encontrados utilizando os modelos de distribuição de créditos ciente de tópicos produzem as maiores propagações (HCD na base do *Flixster* e CD na base do *Epinions*); e (4) o modelo HCD é equiparável ao modelo CD, em relação ao tempo necessário para encontrar os k-usuários do conjunto inicial. Além disso, os dois modelos são ordens de magnitude mais rápidos do que os modelos que utilizam simulações Monte Carlo. Esse fato permite que o conjunto inicial seja reprocessado rapidamente e frequentemente, caso seja necessário.

5.2 Limitações do Trabalho

A seguir, apresenta-se uma visão crítica das limitações deste trabalho, as quais devem ser abordadas para dar continuidade à pesquisa desenvolvida nesta tese.

- **Influência em um único tópico** - A solução apresentada nesta tese foi concebida para encontrar os usuários mais influentes considerando apenas um tópico de interesse. Todavia, em cenários reais pode ser requerido que sejam encontrados usuários influentes em mais de um tópico simultaneamente. Alguns autores, como por exemplo, Barbiere *et al.* [61] e Li *et al.* [101] têm tratado essa questão em trabalhos recentes;

- **Ausência da noção de opinião positiva e negativa** - A solução apresentada considera que os usuários em uma rede social sempre exercem influência social po-

sitiva em relação a um tópico específico. Todavia, alguns autores [42, 68, 102] têm utilizado modelos de propagação que consideram, a presença de usuários com influência social positiva e negativa. Por exemplo, um usuário que propaga muitas informações sobre um tópico específico, pode na verdade estar propagando informações negativas relacionadas àquele tópico. Desse modo, esse usuário não deveria ser selecionado para fazer parte do conjunto inicial;

- **Necessidade de reprocessamento do registro de propagações** - Outra limitação da solução concebida nesta tese, é a necessidade de reprocessar todo o registro de propagações baseado em tópicos para encontrar os usuários mais influentes. Esse problema ocorre devido ao registro de propagações baseado em tópicos representar uma visão estática das ações que estão ocorrendo em tempo real em uma rede social. Desse modo, sempre que novas tuplas forem adicionadas, haverá a necessidade de reprocessar todo o registro de propagações para acomodar as modificações ocorridas. Para aquelas soluções que não são escaláveis, esse problema torna-se ainda mais complexo, uma vez que as mesmas necessitam de muito tempo principalmente para calcular a influência entre os usuários (probabilidades das arestas) para cada tópico. A fim de minimizar esse problema, alguns autores [103] têm proposto soluções que processam novos dados recebidos de forma contínua e, desse modo, não precisam reprocessar todas as tuplas do registro de propagações;

- **Ausência de competição** - Por fim, não faz parte do escopo deste trabalho a noção de ambientes competitivos [64]. No dia a dia, os produtos são manufaturados e revendidos por diversas empresas, as quais estão constantemente disputando novos consumidores com o objetivo de aumentar sua fatia de mercado para aquele produto. Nesse tipo de cenário, onde há a presença de empresas concorrentes, várias delas podem estar interessadas em encontrar simultaneamente usuários que sejam influentes em um determinado tópico. Desse modo, naturalmente elas também irão competir umas com as outras para atrair os mesmos usuários, para fazer com que eles propaguem informações a cerca de seus produtos nas redes sociais, em detrimento dos produtos das empresas rivais.

5.3 Trabalhos Futuros

Além das limitações mencionadas anteriormente, vários rumos de pesquisa podem ser dados com base nas contribuições deste trabalho. Alguns destes potenciais rumos de continuidade são descritos a seguir:

- **Exploração de outros fatores relacionados a Homofilia** - A simples inclusão

dos fatores de gênero e idade na equação de distribuição de créditos diretos produziu um impacto na etapa de seleção do conjunto inicial do modelo HCD (cerca de 20% de novos usuários selecionados). Esse "novo" conjunto inicial foi responsável por um aumento de 3% no tamanho das propagações produzidas pelo conjunto inicial do modelo CD. Dessa forma, observa-se que a escolha dos fatores utilizados nessa equação é, claramente, um aspecto que pode ser melhorado, com o objetivo de obter melhores resultados;

- **Realização de um estudo qualitativo sobre os usuários que fazem parte dos conjuntos iniciais** - Uma questão de pesquisa interessante seria avaliar a presença de determinadas propriedades relacionadas à Análise de Redes Sociais no conjunto inicial de usuários. Esse tipo de estudo ajudaria a compreender quais são as características comuns que os usuários presentes nesses conjuntos iniciais possuem. Uma vez compreendidas, essas características comuns poderiam ser utilizadas para definir papeis para cada tipo de agente presente em uma propagação;

- **Seleção de usuários influentes com base em papeis definidos** - Trata-se de um trabalho complementar ao sugerido no item anterior. Considere, por exemplo, que sejam observados alguns papeis que participam de uma propagação e que sejam conhecidas as proporções desses papeis em cada propagação. Considere ainda que, cada usuário na rede social possa ser classificado em um desses papeis. Em um cenário como esse, seria possível selecionar os usuários de cada papel em uma proporção pré-definida. Essa abordagem poderia ser utilizada para substituir soluções não escaláveis, como aquelas que necessitam determinar as probabilidades das arestas e executar Simulações Monte Carlo.

Referências Bibliográficas

[1] ROGERS, E. M. *Diffusion of Innovations, 5th Edition*. 5th. ed. [S.l.]: Free Press, 2003. ISBN 0743222091.

[2] DAVID, E.; JON, K. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. New York, NY, USA: Cambridge University Press, 2010. ISBN 0521195330, 9780521195331.

[3] WASSERMAN, S.; FAUST, K. *Social Network Analysis: Methods and Applications*. [S.l.]: Cambridge University Press, 1994. (Structural Analysis in the Social Sciences). ISBN 9780521387071.

[4] COLEMAN, J. S.; KATZ, E.; MENZEL, H. *Medical Innovation: A Diffusion Study*. [S.l.]: Bobbs-Merrill Co, 1966. Unknown Binding.

[5] RYAN, B.; GROSS, N. The diffusion of hybrid seed corn in two Iowa communities. *Rural Sociology*, 1943, v. 8, n. 1, p. 15–24, 1943.

[6] SocialTimes. *The Growth of Social Media: From Passing Trend to International Obsession Infographic*. 2013. Disponível em: <`http://socialtimes.com/the-growth-of-social-media-from-trend-to-obsession-infographic_b141318`>. Acesso em: 30 jun. 2015.

[7] THEGUARDIAN. *Facebook: 10 years of social networking, in numbers*. 2014. Disponível em: <`http://www.theguardian.com/news/datablog/2014/feb/04/facebook-in-numbers-statistics`>. Acesso em: 30 jun. 2015.

[8] BOYD, D.; ELLISON, N. Social network sites: definition, history, and scholarship. *IEEE Engineering Management Review*, 2010, v. 38, n. 3, p. 16–31, 2010. ISSN 0360-8581.

[9] KIM, W.; JEONG, O.-R.; LEE, S.-W. On social Web sites. *Information Systems*, 2010, Elsevier Science Ltd., Oxford, UK, UK, v. 35, n. 2, p. 215–236, abr. 2010. ISSN 0306-4379. Disponível em: <http://dx.doi.org/10.1016/j.is.2009.08.003>.

[10] BHAGAT, S.; GOYAL, A.; LAKSHMANAN, L. V. Maximizing Product Adoption in Social Networks. In: *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM, 2012. (WSDM '12), p. 603–612. ISBN 978-1-4503-0747-5. Disponível em: <http://doi.acm.org/10.1145/2124295.2124368>.

[11] SONG, X. et al. Personalized recommendation driven by information flow. In: *Proceedings of the 29th annual international ACM SIGIR conference on*

Research and development in information retrieval. New York, NY, USA: ACM, 2006. (SIGIR '06), p. 509–516. ISBN 1-59593-369-7. Disponível em: <http://doi.acm.org/10.1145/1148170.1148258>.

[12] SONG, X. et al. Information flow modeling based on diffusion rate for prediction and ranking. In: *Proceedings of the 16th international conference on World Wide Web*. New York, NY, USA: ACM, 2007. (WWW '07), p. 191–200. ISBN 978-1-59593-654-7. Disponível em: <http://doi.acm.org/10.1145/1242572.1242599>.

[13] KOREN, Y.; BELL, R.; VOLINSKY, C. Matrix Factorization Techniques for Recommender Systems. *Computer*, 2009, IEEE Computer Society Press, Los Alamitos, CA, USA, v. 42, n. 8, p. 30–37, ago. 2009. ISSN 0018-9162. Disponível em: <http://dx.doi.org/10.1109/MC.2009.263>.

[14] ANAND, S. S.; GRIFFITHS, N. A market-based approach to address the new item problem. In: MOBASHER, B. et al. (Ed.). *RecSys*. [S.l.]: ACM, 2011. p. 205–212. ISBN 978-1-4503-0683-6.

[15] SHANG, S. et al. Wisdom of the Crowd: Incorporating Social Influence in Recommendation Models. In: *Proceedings of the 2011 IEEE 17th International Conference on Parallel and Distributed Systems*. Washington, DC, USA: IEEE Computer Society, 2011. (ICPADS '11), p. 835–840. ISBN 978-0-7695-4576-9. Disponível em: <http://dx.doi.org/10.1109/ICPADS.2011.150>.

[16] MYERS, S. A.; ZHU, C.; LESKOVEC, J. Information diffusion and external influence in networks. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2012. (KDD '12), p. 33–41. ISBN 978-1-4503-1462-6. Disponível em: <http://doi.acm.org/10.1145/2339530.2339540>.

[17] GOYAL, A. et al. GuruMine: A Pattern Mining System for Discovering Leaders and Tribes. In: *Proceedings of the 2009 IEEE International Conference on Data Engineering*. Washington, DC, USA: IEEE Computer Society, 2009. (ICDE '09), p. 1471–1474. ISBN 978-0-7695-3545-6. Disponível em: <http://dx.doi.org/10.1109/ICDE.2009.59>.

[18] WENG, J. et al. TwitterRank: finding topic-sensitive influential twitterers. In: *Proceedings of the third ACM international conference on Web search and data mining*. New York, NY, USA: ACM, 2010. (WSDM '10), p. 261–270. ISBN 978-1-60558-889-6. Disponível em: <http://doi.acm.org/10.1145/1718487.1718520>.

[19] PAL, A.; COUNTS, S. Identifying topical authorities in microblogs. In: *Proceedings of the fourth ACM international conference on Web search and data mining*. New York, NY, USA: ACM, 2011. (WSDM '11), p. 45–54. ISBN 978-1-4503-0493-1. Disponível em: <http://doi.acm.org/10.1145/1935826.1935843>.

[20] SAEZ-TRUMPER, D. et al. Finding trendsetters in information networks. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2012. (KDD '12), p. 1014–1022. ISBN 978-1-4503-1462-6. Disponível em: <http://doi.acm.org/10.1145/2339530.2339691>.

[21] GOYAL, A.; BONCHI, F.; LAKSHMANAN, L. V. Discovering leaders from community actions. In: *Proceedings of the 17th ACM conference on Information and knowledge management*. New York, NY, USA: ACM, 2008. (CIKM '08), p. 499–508. ISBN 978-1-59593-991-3. Disponível em: <http://doi.acm.org/10.1145/1458082.1458149>.

[22] BARBIERI, N.; BONCHI, F.; MANCO, G. Topic-aware social influence propagation models. *Knowledge and Information Systems*, 2013, v. 37, n. 3, p. 555–584, 2013.

[23] RODRIGUEZ, M. G.; LESKOVEC, J.; KRAUSE, A. Inferring networks of diffusion and influence. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2010. (KDD '10), p. 1019–1028. ISBN 978-1-4503-0055-1. Disponível em: <http://doi.acm.org/10.1145/1835804.1835933>.

[24] MATHIOUDAKIS, M. et al. Sparsification of influence networks. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2011. (KDD '11), p. 529–537. ISBN 978-1-4503-0813-7. Disponível em: <http://doi.acm.org/10.1145/2020408.2020492>.

[25] GOMEZ-RODRIGUEZ, M.; LESKOVEC, J.; KRAUSE, A. Inferring Networks of Diffusion and Influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2012, ACM, New York, NY, USA, v. 5, n. 4, p. 21:1–21:37, fev. 2012. ISSN 1556-4681. Disponível em: <http://doi.acm.org/10.1145/2086737.2086741>.

[26] WANG, L.; ERMON, S.; HOPCROFT, J. E. Feature-Enhanced Probabilistic Models for Diffusion Network Inference. In: FLACH, P. A.; BIE, T. D.; CRISTIANINI, N. (Ed.). *ECML/PKDD (2)*. [S.l.]: Springer, 2012. (Lecture Notes in Computer Science, v. 7524), p. 499–514. ISBN 978-3-642-33485-6.

[27] GOLBECK, J.; HENDLER, J. Inferring Binary Trust Relationships in Web-based Social Networks. *ACM Transactions on Internet Technology (TOIT)*, 2006, ACM, New York, NY, USA, v. 6, n. 4, p. 497–529, nov. 2006. ISSN 1533-5399. Disponível em: <http://doi.acm.org/10.1145/1183463.1183470>.

[28] KOTLER, P. *Principles of Marketing*. [S.l.]: Prentice Hall/FinancialTimes, 2005. ISBN 9780273684565.

[29] JANNACH, D. et al. *Recommender Systems: An Introduction*. 1st. ed. New York, NY, USA: Cambridge University Press, 2010. ISBN 0521493366, 9780521493369.

[30] RASHOTTE, L. Social influence. In: *The Blackwell encyclopedia of sociology*. [S.l.: s.n.], 2006. IX, p. 4426–4429.

[31] GRANOVETTER, M. S. The Strength of Weak Ties. *American Journal of Sociology*, 1973, The University of Chicago Press, v. 78, n. 6, p. 1360–1380, 1973. ISSN 00029602. Disponível em: <http://dx.doi.org/10.2307/2776392>.

[32] DOMINGOS, P.; RICHARDSON, M. Mining the network value of customers. In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2001. (KDD '01), p. 57–66. ISBN 1-58113-391-X. Disponível em: <http://doi.acm.org/10.1145/502512.502525>.

[33] RICHARDSON, M.; DOMINGOS, P. Mining knowledge-sharing sites for viral marketing. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2002. (KDD '02), p. 61–70. ISBN 1-58113-567-X. Disponível em: <http://doi.acm.org/10.1145/775047.775057>.

[34] KEMPE, D.; KLEINBERG, J.; TARDOS, E. Maximizing the spread of influence through a social network. In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2003. (KDD '03), p. 137–146. ISBN 1-58113-737-0. Disponível em: <http://doi.acm.org/10.1145/956750.956769>.

[35] GOYAL, A.; BONCHI, F.; LAKSHMANAN, L. V. S. A data-based approach to social influence maximization. *Proceedings of the VLDB Endowment*, 2011, VLDB Endowment, v. 5, n. 1, p. 73–84, set. 2011. ISSN 2150-8097. Disponível em: <http://dl.acm.org/citation.cfm?id=2047485.2047492>.

[36] KIMURA, M.; SAITO, K. Tractable models for information diffusion in social networks. In: *Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases*. Berlin, Heidelberg: Springer-Verlag, 2006. (PKDD'06), p. 259–271. ISBN 3-540-45374-1, 978-3-540-45374-1. Disponível em: <http://dx.doi.org/10.1007/11871637_27>.

[37] LESKOVEC, J. et al. Cost-effective outbreak detection in networks. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2007. (KDD '07), p. 420–429. ISBN 978-1-59593-609-7. Disponível em: <http://doi.acm.org/10.1145/1281192.1281239>.

[38] CHEN, W.; WANG, Y.; YANG, S. Efficient influence maximization in social networks. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2009. (KDD '09), p. 199–208. ISBN 978-1-60558-495-9. Disponível em: <http://doi.acm.org/10.1145/1557019.1557047>.

[39] CHEN, W.; YUAN, Y.; ZHANG, L. Scalable Influence Maximization in Social Networks under the Linear Threshold Model. In: *Proceedings of the 2010 IEEE International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2010. (ICDM '10), p. 88–97. ISBN 978-0-7695-4256-0. Disponível em: <http://dx.doi.org/10.1109/ICDM.2010.118>.

[40] WANG, Y. et al. Community-based greedy algorithm for mining top-K influential nodes in mobile social networks. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2010. (KDD '10), p. 1039–1048. ISBN 978-1-4503-0055-1. Disponível em: <http://doi.acm.org/10.1145/1835804.1835935>.

[41] GOYAL, A.; LU, W.; LAKSHMANAN, L. V. CELF++: optimizing the greedy algorithm for influence maximization in social networks. In: *Proceedings of the 20th international conference companion on World wide web*. New York, NY, USA: ACM, 2011. (WWW '11), p. 47–48. ISBN 978-1-4503-0637-9. Disponível em: <http://doi.acm.org/10.1145/1963192.1963217>.

[42] LI, Y. et al. Influence Diffusion Dynamics and Influence Maximization in Social Networks with Friend and Foe Relationships. In: *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM, 2013. (WSDM '13), p. 657–666. ISBN 978-1-4503-1869-3. Disponível em: <http://doi.acm.org/10.1145/2433396.2433478>.

[43] BORGS, C. et al. Maximizing Social Influence in Nearly Optimal Time. In: *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2014. (SODA '14), p. 946–957. ISBN 978-1-611973-38-9. Disponível em: <http://dl.acm.org/citation.cfm?id=2634074.2634144>.

[44] TANG, Y.; XIAO, X.; SHI, Y. Influence Maximization: Near-optimal Time Complexity Meets Practical Efficiency. In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA: ACM, 2014. (SIGMOD '14), p. 75–86. ISBN 978-1-4503-2376-5. Disponível em: <http://doi.acm.org/10.1145/2588555.2593670>.

[45] CHEN, W.; WANG, C.; WANG, Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2010. (KDD '10), p. 1029–1038. ISBN 978-1-4503-0055-1. Disponível em: <http://doi.acm.org/10.1145/1835804.1835934>.

[46] GOYAL, A.; LU, W.; LAKSHMANAN, L. V. S. SIMPATH: An Efficient Algorithm for Influence Maximization under the Linear Threshold Model. In: *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2011. (ICDM '11), p. 211–220. ISBN 978-0-7695-4408-3. Disponível em: <http://dx.doi.org/10.1109/ICDM.2011.132>.

[47] JIANG, Q. et al. Simulated Annealing Based Influence Maximization in Social Networks. In: *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*. AAAI Press, 2011. p. 127–132. Disponível em: <https://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3670/3841>.

[48] CHEN, Y.-C.; PENG, W.-C.; LEE, S.-Y. Efficient algorithms for influence maximization in social networks. *Knowledge and Information Systems*, 2012, v. 33, n. 3, p. 577–601, 2012.

[49] JUNG, K.; HEO, W.; CHEN, W. IRIE: Scalable and Robust Influence Maximization in Social Networks. In: *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2012. (ICDM '12), p. 918–923. ISBN 978-0-7695-4905-7. Disponível em: <http://dx.doi.org/10.1109/ICDM.2012.79>.

[50] WANG, C.; CHEN, W.; WANG, Y. Scalable influence maximization for independent cascade model in large-scale social networks. *Data Mining and Knowledge Discovery*, 2012, Springer US, v. 25, n. 3, p. 545–576, 2012. ISSN 1384-5810. Disponível em: <http://dx.doi.org/10.1007/s10618-012-0262-1>.

[51] CHENG, S. et al. IMRank: Influence Maximization via Finding Self-consistent Ranking. In: *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. New York, NY, USA: ACM, 2014. (SIGIR '14), p. 475–484. ISBN 978-1-4503-2257-7. Disponível em: <http://doi.acm.org/10.1145/2600428.2609592>.

[52] BENEVENUTO, F.; ALMEIDA, J.; SILVA, A. Coleta e Análise de Grandes Bases de Dados de Redes Sociais Online. In: *Jornadas de Atualização em Informática (JAI)*. [S.l.: s.n.], 2011. p. 11–57.

[53] SAITO, K.; NAKANO, R.; KIMURA, M. Prediction of Information Diffusion Probabilities for Independent Cascade Model. In: *Proceedings of the 12th international conference on Knowledge-Based Intelligent Information and Engineering Systems, Part III*. Berlin, Heidelberg: Springer-Verlag, 2008. (KES '08), p. 67–75. ISBN 978-3-540-85566-8. Disponível em: <http://dx.doi.org/10.1007/978-3-540-85567-5_9>.

[54] SAITO, K. et al. Learning diffusion probability based on node attributes in social networks. In: *Proceedings of the 19th international conference on Foundations of intelligent systems*. Berlin, Heidelberg: Springer-Verlag, 2011. (ISMIS'11), p. 153–162. ISBN 978-3-642-21915-3. Disponível em: <http://dl.acm.org/citation.cfm?id=2029759.2029781>.

[55] GOYAL, A.; BONCHI, F.; LAKSHMANAN, L. V. Learning influence probabilities in social networks. In: *Proceedings of the third ACM international conference on Web search and data mining*. New York, NY, USA: ACM, 2010. (WSDM '10), p. 241–250. ISBN 978-1-60558-889-6. Disponível em: <http://doi.acm.org/10.1145/1718487.1718518>.

[56] BARBIERI, N.; BONCHI, F.; MANCO, G. Cascade-based community detection. In: *Proceedings of the sixth ACM international conference on Web search and data mining*. New York, NY, USA: ACM, 2013. (WSDM '13), p. 33–42. ISBN 978-1-4503-1869-3. Disponível em: <http://doi.acm.org/10.1145/2433396.2433403>.

[57] GRANOVETTER, M. Threshold Models of Collective Behavior. *American Journal of Sociology*, 1978, The University of Chicago Press, v. 83, n. 6, p. 1420–1443, 1978. ISSN 00029602. Disponível em: <http://dx.doi.org/10.2307/2778111>.

[58] LIU, L. et al. Mining topic-level influence in heterogeneous networks. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. New York, NY, USA: ACM, 2010. (CIKM '10), p. 199–208. ISBN 978-1-4503-0099-5. Disponível em: <http://doi.acm.org/10.1145/1871437.1871467>.

[59] TANG, J. et al. Social influence analysis in large-scale networks. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2009. (KDD '09), p. 807–816. ISBN 978-1-60558-495-9. Disponível em: <http://doi.acm.org/10.1145/1557019.1557108>.

[60] ZHANG, Y.; ZHOU, J.; CHENG, J. Preference-Based Top-K Influential Nodes Mining in Social Networks. In: *Proceedings of the 2011 IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications*. Washington, DC, USA: IEEE Computer Society, 2011. (TRUSTCOM '11), p. 1512–1518. ISBN 978-0-7695-4600-1. Disponível em: <http://dx.doi.org/10.1109/TrustCom.2011.209>.

[61] ASLAY, Ç. et al. Online topic-aware influence maximization queries. In: *Proceedings of the 17th International Conference on Extending Database Technology, EDBT 2014, Athens, Greece, March 24-28, 2014.* [s.n.], 2014. p. 295–306. Disponível em: <http://dx.doi.org/10.5441/002/edbt.2014.28>.

[62] CHEN, W.; LIN, T.; YANG, C. Efficient topic-aware influence maximization using preprocessing. *CoRR*, 2014, abs/1403.0057, 2014. Disponível em: <http://arxiv.org/abs/1403.0057>.

[63] CHEN, S. et al. Online Topic-aware Influence Maximization. *Proceedings of the VLDB Endowment*, 2015, VLDB Endowment, v. 8, n. 6, p. 666–677, fev. 2015. ISSN 2150-8097. Disponível em: <http://dx.doi.org/10.14778/2735703.2735706>.

[64] LI, H. et al. GetReal: Towards Realistic Selection of Influence Maximization Strategies in Competitive Networks. In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA: ACM, 2015. (SIGMOD '15), p. 1525–1537. ISBN 978-1-4503-2758-9. Disponível em: <http://doi.acm.org/10.1145/2723372.2723710>.

[65] MCPHERSON, M.; SMITH-LOVIN, L.; COOK, J. M. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 2001, v. 27, n. 1, p. 415–444, 2001. Disponível em: <http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.soc.27.1.415>.

[66] KIMURA, M.; SAITO, K.; MOTODA, H. Blocking links to minimize contamination spread in a social network. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2009, ACM, New York, NY, USA, v. 3, n. 2, p. 9:1–9:23, abr. 2009. ISSN 1556-4681. Disponível em: <http://doi.acm.org/10.1145/1514888.1514892>.

[67] BUDAK, C.; AGRAWAL, D.; ABBADI, A. E. Structural trend analysis for online social networks. *Proceedings of the VLDB Endowment*, 2011, VLDB Endowment, v. 4, n. 10, p. 646–656, jul. 2011. ISSN 2150-8097. Disponível em: <http://dl.acm.org/citation.cfm?id=2021017.2021022>.

[68] BUDAK, C.; AGRAWAL, D.; ABBADI, A. E. Limiting the spread of misinformation in social networks. In: *Proceedings of the 20th international conference on World wide web*. New York, NY, USA: ACM, 2011. (WWW '11), p. 665–674. ISBN 978-1-4503-0632-4. Disponível em: <http://doi.acm.org/10.1145/1963405.1963499>.

[69] HE, X. et al. Influence Blocking Maximization in Social Networks under the Competitive Linear Threshold Model. In: *SDM'12*. [S.l.: s.n.], 2012. p. 463–474.

[70] DEGENNE, A.; FORSÉ, M. *Introducing Social Networks*. SAGE Publications, 1999. (ISM (London, England)). ISBN 9780761956044. Disponível em: <http://books.google.com.br/books?id=D2_LW66BRgoC>.

[71] SZWARCFITER, J. *Grafos e algoritmos computacionais*. [S.l.]: Campus, 1984.

[72] RECUERO, R. *Redes sociais na internet*. Sulina, 2009. (Coleção cibercultura). ISBN 9788520505250. Disponível em: <http://books.google.com.br/books?id=yQaLPgAACAAJ>.

[73] STEGLICH, C.; SNIJDERS, T. A. B.; PEARSON, M. *Dynamic Networks and Behavior: Separating Selection from Influence*. [S.l.], Dez 2006. Whitepaper. Disponível em: <http://www.ppsw.rug.nl/ steglich/pdf/SSPrevised.pdf>.

[74] ANAGNOSTOPOULOS, A.; KUMAR, R.; MAHDIAN, M. Influence and Correlation in Social Networks. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2008. (KDD '08), p. 7–15. ISBN 978-1-60558-193-4. Disponível em: <http://doi.acm.org/10.1145/1401890.1401897>.

[75] CRANDALL, D. et al. Feedback Effects Between Similarity and Social Influence in Online Communities. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2008. (KDD '08), p. 160–168. ISBN 978-1-60558-193-4. Disponível em: <http://doi.acm.org/10.1145/1401890.1401914>.

[76] ARAL, S.; MUCHNIK, L.; SUNDARARAJAN, A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 2009, National Academy of Sciences, v. 106, n. 51, p. 21544–21549, dez. 2009. ISSN 1091-6490. Disponível em: <http://dx.doi.org/10.1073/pnas.0908800106>.

[77] FOND, T. L.; NEVILLE, J. Randomization Tests for Distinguishing Social Influence and Homophily Effects. In: *Proceedings of the 19th International Conference on World Wide Web*. New York, NY, USA: ACM, 2010. (WWW '10), p. 601–610. ISBN 978-1-60558-799-8. Disponível em: <http://doi.acm.org/10.1145/1772690.1772752>.

[78] KWAK, H. et al. What is Twitter, a Social Network or a News Media? In: *Proceedings of the 19th International Conference on World Wide Web*. New York, NY, USA: ACM, 2010. (WWW '10), p. 591–600. ISBN 978-1-60558-799-8. Disponível em: <http://doi.acm.org/10.1145/1772690.1772751>.

[79] CHA, M. et al. Measuring User Influence in Twitter: The Million Follower Fallacy. In: *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*. Washington DC, USA: [s.n.], 2010. p. 10–17.

[80] SILVA, A. et al. ProfileRank: Finding Relevant Content and Influential Users Based on Information Diffusion. In: *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. New York, NY, USA: ACM, 2013. (SNAKDD '13), p. 1–9. ISBN 978-1-4503-2330-7. Disponível em: <http://doi.acm.org/10.1145/2501025.2501033>.

[81] BRIN, S.; PAGE, L. The Anatomy of a Large-scale Hypertextual Web Search Engine. *Comput. Netw. ISDN Syst.*, 1998, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 30, n. 1-7, p. 107–117, abr. 1998. ISSN 0169-7552. Disponível em: <http://dx.doi.org/10.1016/S0169-7552(98)00110-X>.

[82] KLEINBERG, J. M. Authoritative Sources in a Hyperlinked Environment. In: *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1998. (SODA '98), p. 668–677. ISBN 0-89871-410-9. Disponível em: <http://dl.acm.org/citation.cfm?id=314613.315045>.

[83] BAKSHY, E. et al. Everyone's an Influencer: Quantifying Influence on Twitter. In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM, 2011. (WSDM '11), p. 65–74. ISBN 978-1-4503-0493-1. Disponível em: <http://doi.acm.org/10.1145/1935826.1935845>.

[84] SUN, J.; TANG, J. A Survey of Models and Algorithms for Social Influence Analysis. In: AGGARWAL, C. C. (Ed.). *Social Network Data Analytics*. Springer US, 2011. p. 177–214. ISBN 978-1-4419-8461-6. Disponível em: <http://dx.doi.org/10.1007/978-1-4419-8462-3_7>.

[85] FOWLER, J. H.; CHRISTAKIS, N. A. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *BMJ*, 2008, British Medical Journal Publishing Group, v. 337, p. 1–9, 2008. ISSN 1468-5833.

[86] KEMPE, D.; KLEINBERG, J.; TARDOS, E. Influential nodes in a diffusion model for social networks. In: *Proceedings of the 32nd international conference on Automata, Languages and Programming*. Berlin, Heidelberg: Springer-Verlag, 2005. (ICALP'05), p. 1127–1138. ISBN 3-540-27580-0, 978-3-540-27580-0. Disponível em: <http://dx.doi.org/10.1007/11523468_91>.

[87] FIGUEIREDO, C. M. H.; MARROQUIM, R. *Teoria dos Grafos*. 2012. Disponível em: <`http://www.cos.ufrj.br/~marroquim/grafos/`>. Acesso em: 30 jun. 2015.

[88] FIGUEIREDO, J. C. A. *Teoria dos Grafos*. 2003. Disponível em: <`http://www.dsc.ufcg.edu.br/~abrantes/CursosAnteriores/tg032.html`>. Acesso em: 30 jun. 2015.

[89] MARIANI, A. C. *Teoria dos Grafos*. Disponível em: <`http://www.inf.ufsc.br/grafos/livro.html`>. Acesso em: 30 jun. 2015.

[90] WIKIPEDIA. *Teoria dos Grafos*. 2015. Disponível em: <`https://pt.wikipedia.org/wiki/Teoria_dos_grafos`>. Acesso em: 30 jun. 2015.

[91] CARVALHO, M. A. G. *Teoria dos Grafos - Uma Introdução*. 2005. Disponível em: <`http://www.ft.unicamp.br/~magic/ft024/apografos_ceset_magic.pdf`>. Acesso em: 30 Jun. 2015.

[92] GOLDBARG, M. C.; GOLDBARG, E. *Grafos: conceitos, algoritmos e aplicações*. [S.l.]: Elsevier, 2012. ISBN 9788535257168.

[93] PRESTES, E. *Graph Theory*. 2011. Disponível em: <`http://www.inf.ufrgs.br/~prestes/Courses/Graph%20Theory/`>. Acesso em: 30 jun. 2015.

[94] CORMEN, T. H. et al. *Introduction to Algorithms*. Second. [S.l.]: MIT Press, 2001.

[95] MENEZES, P. B. *Matemática Discreta para Computação e Informática*. 3. ed. Porto Alegre: Bookman, 2010. (Série didático informática UFRGS).

[96] BLEI, D. M. Probabilistic topic models. *Communications of the ACM*, 2012, ACM, New York, NY, USA, v. 55, n. 4, p. 77–84, abr. 2012. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/2133806.2133826>.

[97] MASSA, P.; AVESANI, P. Trust-aware bootstrapping of recommender systems. In: *ECAI 2006 Workshop on Recommender Systems*. [S.l.: s.n.], 2006. p. 29–33.

[98] JAMALI, M.; ESTER, M. A matrix factorization technique with trust propagation for recommendation in social networks. In: *Proceedings of the fourth ACM conference on Recommender systems*. New York, NY, USA: ACM, 2010. (RecSys '10), p. 135–142. ISBN 978-1-60558-906-0. Disponível em: <http://doi.acm.org/10.1145/1864708.1864736>.

[99] HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123814790, 9780123814791.

[100]WITTEN, I. H.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005. ISBN 0120884070.

[101]LI, Y.; ZHANG, D.; TAN, K.-L. Real-time Targeted Influence Maximization for Online Advertisements. *Proceedings of the VLDB Endowment*, 2015, VLDB Endowment, v. 8, n. 10, p. 1070–1081, jun. 2015. ISSN 2150-8097. Disponível em: <http://dl.acm.org/citation.cfm?id=2794367.2794376>.

[102]CHEN, W. et al. Influence Maximization in Social Networks When Negative Opinions May Emerge and Propagate. In: *SDM*. [S.l.]: SIAM / Omnipress, 2011. p. 379–390. ISBN 978-0-898719-92-5.

[103]KUTZKOV, K. et al. STRIP: Stream Learning of Influence Probabilities. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2013. (KDD '13), p. 275–283. ISBN 978-1-4503-2174-7. Disponível em: <http://doi.acm.org/10.1145/2487575.2487657>.