UNIVERSIDADE FEDERAL DE SÃO JOÃO DEL-REI

Thiago Schons

Análise de Padrões em Espalhamento Gerado por Modelos

São João del-Rei

2015

UNIVERSIDADE FEDERAL DE SÃO JOÃO DEL-REI

Thiago Schons

Análise de Padrões em Espalhamento Gerado por Modelos

Monografia apresentada como requisito da disciplina de Projeto Orientado em Computação II do Curso de Bacharelado em Ciência da Computação da UFSJ.

Orientador: Vinicius Vieira

Universidade Federal de São João del-Rei – UFSJ

Bacharelado em Ciência da Computação

São João del-Rei

2015

Thiago Schons

Análise de Padrões em Espalhamento Gerado por Modelos

Monografia apresentada como requisito da disciplina de Projeto Orientado em Computação II do Curso de Bacharelado em Ciência da Computação da UFSJ.

Trabalho aprovado. São João del-Rei, 18 de julho de 2014:

Vinicius Vieira
Orientador

Professor
Convidado 1

Professor
Convidado 2

São João del-Rei
2015

Este trabalho é dedicado todos que participaram da minha formação, principalmente meus familiares, amigos e professores.

Agradecimentos

Primeiramente, quero agradecer meus pais, Atalíbio e Romilda, que apesar da distância me apoiam incondicionalmente. Meus irmãos, Rosane e Roberto que com suas famílias permitiram que este sonho se tornasse realidade. Vocês todos são o meu alicerce, tenho muito orgulho de poder fazer parte dessa família. Meus amigos, que presentes ou distantes, de Santa Catarina, Belo Horizonte ou de São João, estiveram ao meu lado nas horas boas e difíceis. Gostaria também de mandar um agradecimento ao pessoal da minha república, que tornou mais fácil essa jornada desde a época da pensão. E a todas as pessoas que participaram dessa fase, que com certeza foi uma das melhores da minha vida, e rendeu histórias que contarei aos meus netos.

O sonho é uma fonte infinita de inspiração.

- Luiz Tambucci

Resumo

Atualmente a tecnologia vem propiciando o acesso a redes sociais em toda a sociedade e este acesso vem crescendo de forma significativa. Deste modo, grandes quantidades de interações entre usuários nos mais diversos tipos de redes podem ser mapeadas. O presente trabalho busca investigar a aplicabilidade de uma abordagem para identificação de padrões recorrentes de espalhamento de informações nas mais diferentes formas utilizando algumas métricas já abordadas na literatura. A abordagem de espalhamento utiliza um algoritmo que simula a difusão de boatos na rede que é o Independent Cascade Model, para verificar como é feito o espalhamento variando a probabilidade de um vértice influenciar outro, isso é feito utilizando duas métricas de escolha de probabilidade. As escolhas dos nós iniciais, ou que lançam o boato na rede é feita com os vértices que tem maior grau. A partir do espalhamento gerado é aplicado um algoritmo chamado *Assign Canonical Names* que encontra o nome canônico das árvores geradas buscando os padrões que são mais frequentes na rede, isto é feito para diversas redes, com o intuito de encontrar padrões que sejam comúns nas mais diversas redes.

Palavras-chaves: Redes Complexas, Nomes Canônicos, Espalhamento em Redes, Redes Sociais.

Abstract

Nowadays, technology has been providing access to social networks through society and this access has been growing significantly. This way, big amounts of interactions between users in various types of networks can be mapped. This present work seeks to investigate the applicability of an approach to identify recurrent patterns of information spreading in the most varied forms using some metrics already used in literature. The approach of spreading uses an algorithm that simulates rumor spreading in the "Independent Cascade Model" network, to verify how the spreading is created varying the probability of a vertex influencing another. That is done by using two metrics of probability choosing. The choice of the initial vertexes or that start a rumor on the network is done with the vertexes with bigger degree. From the generated spreading an algorithm called "Assign Canonical Names" is applied, and it finds the canonical name of the generated trees searching for the most frequent patterns on the network. That is done for numerous networks, with the aim of finding patterns that are common for most of them.

Key-words: Complex Networks, Spreading, Canonical Names, Social Networks

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Uma rede em sua forma simples é uma coleção de pontos ligados par a par por linhas, estes pontos são denominados vértices e as linhas arestas. O estudo de redes é muito importante pois a modelagem de fenômenos e processos pode ser feita utilizando redes, tornando esta área muito vasta e abrangente. Qualquer cenário onde há relação entre elementos pode ser modelado como uma rede, e o interesse de estudo por este motivo é imenso. Áreas como biologia, sociologia, física, ciência da computação, matemática e outras, fazem grande uso da modelagem através de redes, resultando assim em grandes benefícios para a área, devido à grande quantidade de pesquisadores atuando e contribuindo.

A modelagem é um dos principais motivos do avanço nas pesquisas da área de redes, visto que há grande aplicabilidade na modelagem de fenomenos do cotidiano, seja interação entre pessoas, animais, objetos, bactérias, enzimas, e diversos outros, gerando uma aplicabilidade imensa com muito interesse nas mais diversas linhas de pesquisa (SCHMITH et al., 2005; ANDERSON et al., 2012). A partir dos dados modelados, utilizando como exemplo a internet, temos que nesta rede os vértices são os computadores ou dispositivos computacionais conectados na rede, e as conexões físicas entre os vértices, fibra óptica, cabos de rede, linhas telefônicas ou sinais de rádio são as arestas.

O relacionamento entre os vértices da rede são a parte de maior interesse do estudo, no qual o objetivo é compreender o motivo das interações entre os vértices, porque eles vieram a ter alguma ligação, e qual o significado de tudo isso, constituindo assim a motivação deste trabalho.

Um exemplo de uma rede intuitiva, a qual utilizamos constantemente, é a *web*. Neste caso os vértices são as páginas e as arestas são os *hyperlinks* contidos entre páginas. Tomando os vértices como pessoas e as arestas como alguma relação que dois vértices podem ter, como por exemplo amizade, namoro, etc, temos a definição de uma rede social, que são de grande interesse em áreas como a sociologia, da qual já possui uma longa tradição de estudos das redes sociais do mundo real, a partir dos quais desenvolveu-se muitas ferramentas matemáticas e estatísticas.

Um dos temas mais importantes no estudo de redes sociais está relacionado à propagação de informação. Um boato, por exemplo, pode ser espalhado em uma rede seguindo alguns padrões, mas tem características únicas: pode atingir muitas ou poucas pessoas, influenciar diversas a também espalharem o boato, entre outras. Com isso, pode-se explorar diversos fatores que permitem compreender como esse boato será propagado, principalmente a influência que o nó exerce sobre seus adjacentes, aumentando assim a

chance de algum vértice adjacente aderir ao boato e prosseguir espalhando-o pelo grafo, podendo aumentar assim a profundidade do mesmo e o grau de aderência do conteúdo.

O cascateamento de informações em uma rede social é de suma importância e pode ser explorado de diversas maneiras. Por exemplo, pode-se difundir algum conteúdo publiciário, identificando nós que são influentes em determinado assunto e maximizando a quantidade de pessoas que terão acesso ao conteúdo.

Neste contexto, este trabalho propõe uma análise sobre dados de redes sociais a fim de verificar como é feita a difusão de um conteúdo, quais padrões são formados na rede, como ocorre uma cascata de informações e que tipos de usuários influenciam na sua difusão. A variação de parâmetros de um modelo de difusão é utilizada para tentar compreender as causas de sucesso do espalhamento de um boato e investigar quais tipos de padrões são mais comuns e em que cenários eles mais ocorrem. Diversos modelos de espalhamento podem ser encontrados na literatura, tais como: *Linear Threshold Model, Independent Cascade Model* e outros (WU; WANG, 2014). Neste trabalho, será utilizado o *Independent Cascade Model* (WANG; CHEN; WANG, 2012) e a análise será feita com base nos resultados obtidos pela aplicação do modelo a um conjunto de redes reais.

1.1 Objetivos Gerais

Normalmente os modelos de propagação são utilizados para análise da formação de cascatas e contagem do número de indivíduos atingidos e a verificação se usuários-alvo foram atingidos. Por outro lado, buscou-se investigar não apenas quantos indivíduos foram atingidos, mas também como ocorreu o espalhamento. Um estudo exploratório com uma abordagem nessa linha foi feito por (DOW; ADAMIC; FRIGGERI, 2013) comparando dois memes lançados na internet. Outro estudo semelhante foi feito por (GOEL; WATTS; GOLDSTEIN, 2012), que analisou características de cascatas reais. O presente trabalho segue a linha de pesquisa de (GOEL; WATTS; GOLDSTEIN, 2012) realizando um estudo semelhante, porém utilizando redes reais e aplicando modelos de difusão para identificar o impacto da variação de parâmetros de um modelo de difusão nas cascatas. Assim mais do que identificar quantos indivíduos são atingidos quando um determinado modelo com determinados parâmetros também são investigados quais são os padrões mais frequentes em diversos tipos de redes e como são as cascatas geradas.

1.2 Objetivos Específicos

O objetivo específico deste trabalho tem como intuito cumprir algumas etapas:

- Desenvolver algoritmo que irá efetuar os testes nas bases de dados;

- Analisar as cascatas formadas pelo espalhamento gerado com o modelo proposto;
- Buscar padrões recorrentes nos mais diversos tipos de redes verificando se há algum ou alguns padrões frequentes identificados em maioria delas e identificando a proporção com que se repetem;
- Verificar os cenários em que ocorrem maior cascateamento de informações;

1.3 Justificativa

Atualmente com a quantidade imensa de informações que são geradas na internet, principalmente em redes sociais, são disponibilizados muitos dados que poderiam gerar um conhecimento de extrema importância se fossem analisados.

A análise de como é feita a propagação e quais parâmetros influenciam na difusão de conteúdos na rede pode colaborar muito para técnicas de difusão de produtos, notícias e informações destinadas à um determinado nicho de consumidores, ou seja, que deva chegar até um determinado público alvo.

O restante do trabalho se organiza da seguinte forma. No capítulo 2 são apresentados os conceitos necessários para entendimento do trabalho realizado e dos capítulos seguintes. No capítulo 3 são descritos alguns trabalhos que se relacionam com a metodologia proposta e são base do presente trabalho. No capítulo 4 é mostrada a metodologia utilizada. No capítulo 5 é descrita a implementação dos algoritmos com os pseudocódigos e descrição dos mesmos. No capítulo 6 é feita uma discussão dos resultados e no capítulo 7 é dada uma conclusão com propostas para trabalhos futuros.

2 Referencial Teórico

Neste capítulo será descrito de maneira mais geral alguns tópicos importantes em relação ao presente trabalho. Em razão de uma boa compreensão do estudo, serão fornecidos alguns conceitos referentes à Redes Sociais, Redes Complexas e Espalhamento em Redes, assuntos aos quais este trabalho está relacionado.

2.1 Redes Complexas

Uma rede complexa nada mais é do que um grafo e tem em sua representação uma coleção de pontos (vértices) conectados por linhas (arestas). O conjunto destes nós e arestas denomina-se uma rede(NEWMAN, 2010). A rede nos deixa modelar relações à partir de diferentes contextos e aplica-se a muitas áreas da ciência, como física, biologia, ciências sociais, neurociência, psicologia, redes de comunicação e outras.

A natureza das partes individuais da rede são muito estudadas, principalmente de onde surgem e porque surgem as interações entre as redes e o que pode correlacioná-las. Temos como exemplo redes de computadores, que podem ser modeladas com grafos. Podemos até modelar como um ser humano age, tentando identificar o motivo de locomoção do mesmo(SCHNEIDER et al., 2013), e outras tantas propriedades. A busca por padrões entre estas redes é algo que está sendo muito buscado nos últimos tempos, visto que assim pode-se mapear muitas coisas afim de identificar razões para determinados padrões estarem ocorrendo. Os padrões e a estrutura da rede tem grande efeito no comportamento que a rede terá. A internet por exemplo, depende dos seus usuários o resultado do roteamento escolhido para determinado pacote de dados chegar. Em uma rede social, as conexões afetam como as pessoas aprendem, formam opiniões, lêem notícias, etc.

Para representar uma rede, muitos dados da mesma são perdidos, isso é feito para tornar possível a modelagem, gerando como consequencia muitas desvantagens, por inutilizar muitos dados que poderiam ser importantes, porém também criando vantagens, principalmente por tornar possível a modelagem e deixar a representação mais fácil de ser visualizada e entendida.

Hoje em dia, temos uma grande variedade de ferramentas e algoritmos para analisar, modelar e entender redes, muitas propriedades importantes são tiradas disso, como uma simples representação e posteriormente, através de cálculos podemos verificar os vértices mais importantes da rede, o grupos de vértices melhor conectado, melhores caminhos para transmissão e outras diversas propriedades.

2.1.1 Propriedades das Redes

Se já temos os dados de alguma rede o que fazer com eles? Primeiramente devemos modelar, posteriormente uma análise deve ser feita. Atualmente existem diversas métricas que foram desenvolvidas para análise de redes, esse desenvolvimento foi feito através da necessidade de entender redes muito grandes, onde humanos não conseguem interpretar, tudo isso para nos ajudar entender o que ocorre e o que a rede quer nos dizer. Mesmo em casos onde a visualização é impossível.

Alguns conceitos de redes são introduzidos rapidamente e de forma simples, como observados em (GHOSHAL, 2011) e podem ser de suma importância. Um conceito importante é o de centralidade, a centralidade diz o quão importante vértices ou arestas são na rede. Outra medida é o grau de um vértice, que pode significar muito, vértices com grau alto se ligam a muitas pessoas e grau baixo a poucas pessoas. Dependendo da rede, o grau de um vertice e a quantidade de arestas que incidem nele, ainda podemos filtrar isto em grau de entrada "In-degree" e grau de saída "out-degree" que no caso de um grafo direcionado, onde as arestas saem de um vértice e apontam para outro, são os vértices que são apontados e que apontam, respectivamente.

Algumas aglomerações nas redes são importantes para medir o fenômeno de transitividade que o grafo tem, ou seja, o quanto os vértices formam "triângulos" conectando um vértice A ao vértice B e um vértice B ao vértice C, aumentando assim as chances do vértice A estar conectado ao vértice C, medindo a redundância das arestas ao redor de um vértice.

Outro exemplo de um conceito de rede que surge várias vezes e tem implicações práticas reais é o chamado efeito pequeno mundo, ou "small-world", também conhecido como seis graus de separação, onde este fenômeno diz que se forem escolhidas duas pessoas quaisquer na Terra, pode-se encontrar um caminho de até seis outras pessoas conhecidas entre elas, esta é uma das razões da internet funcionar atualmente, pois atinge computadores em poucos passos, transmitindo dados mais rapidamente (GRANOVETTER, 1978).

2.2 Espalhamento em Redes

O espalhamento em redes é aplicado em diversas áreas, e há um interesse imenso em pesquisas na área, principalmente por ter surgido relativamente há pouco tempo, todo e qualquer esforço poderá colaborar muito para o desenvolvimento da mesma. Segundo (WU; WANG, 2014), a influência de difusão é o processo que a informação adquire para se propagar através de indivíduos em uma rede social. O início dos estudos de difusão foram dados no século 20. O primeiro a descrever um modelo matemático formal no assunto foi (GRANOVETTER, 1978). Hoje temos muitos modelos que são derivados de diversas

áreas, e alguns deles são mais utilizados tais como os modelos Threshold Linear e o Cascading Independent Model, que são mais comumente utilizados para influências sociais, porém tem também modelos voltados para redes biológicas, utilizados para propagação de doenças.

Atualmente o investimento na área de redes é imenso, e isso se deve ao fato da modelagem de redes ser atribuída aos mais variados cenários. Em redes sociais e no contágio de doenças, seja pelo contato, por estar em algum ambiente com uma pessoa doente, ou até o HIV que normalmente é transmitido quando duas pessoas tem contato sexual sem proteção. A estrutura de redes, seja na modelagem ou nos algoritmos para resolverem seus problemas, se encaixa perfeitamente neste ambiente e em tantos outros, que são estudados nas mais diversas áreas, seja computacional, biológica, médica, e até social.

Atualmente temos diversos problemas com vírus e disseminação de informações na rede, até porque temos os "vírus" computacionais que se reproduzem automaticamente na rede, através de *downloads* e acessos feitos por usuários, eles se infectam em dispositivos computacionais da mesma forma que se propagam infecções patogênicas entre pessoas e animais. Porém, isto pode ser modelado através de redes.

2.2.1 Modelos de Propagação

Alguns modelos foram desenvolvidos na área de biologia quando um "portador" pega uma infecção e tende a disseminá-la de alguma forma a alguém próximo ou que tenha algum tipo de contato com o mesmo. A seguir serão descritos 3 modelos: SI, SIS e SIS, além do complemento destes modelos que são SIS e SIRS, o embasamento teórico dos modelos epidemiológicos foram extraídos de (NEWMAN, 2010) e (DODDS; WATTS, 2005), no caso os modelos são: SI, SIR, SIS e SIRS.

Temos também vários modelos em cascata, no qual está inserido o *Independent Cascade Model* que é o modelo adotado para execução dos testes, a explicação dele se dará na subseção 2.2.2.

2.2.1.1 Modelo SI

Na típica representação matemática de uma epidemia dentro de um hospedeiro, a doença é resumida em alguns estados básicos. O modelo SI tem a versão mais resumida destes modelos, que contém somente dois estados, que é suscetível e infectado. O primeiro estado (suscetível), ocorre quando há um indivíduo que não é portador da doença ainda, porém ao entrar em contato com algum portador, pode vir a ser um novo portador. E o estado "infectado" ocorre quando o indivíduo está com a doença e pode assim propagá-la para outros indivíduos que estejam no estado "suscetível" dos quais venham a ter algum tipo de contato.

Embora este modelo deixe uma grande quantidade de detalhes escondidos, que poderiam ser usados, ele capta muitas características da dinâmica de doenças que podem ser simplificadamente utilizadas em casos onde queremos saber o que está acontecendo, principalmente no nível de redes e populações para ver como a propagação das doenças se comportam. Este modelo é muito bem utilizado quando o indivíduo infectado não volta a ser suscetível, ou seja, não há recuperação da infecção.

2.2.1.2 Modelo SIR

O modelo SIR também é um modelo muito simples de infecção, nele há muitas maneiras das quais este modelo pode ser estendido e tornar-se mais realista ou apropriado para alguma doença específica. Nos indivíduos do modelo SI, uma vez infectado, está infectado eternamente, porém na realidade, após um certo tempo a infecção pode passar, pois o corpo do indivíduo combate a doença através de seu sistema imunológico, e além disso, ao adquirir anticorpos a chance de obter a doença novamente é menor. Para representar este comportamento são necessários 3 estados: suscetível, infectado e recuperado, este modelo é o modelo SIR.

2.2.1.3 Modelos SIS e SIRS

Os Modelo SIS e SIRS são os mesmos modelos que o SI e o SIR, respectivamente, com uma única diferença, ambos tem a possibilidade de reinfecção, ou seja, se um indivíduo for contaminado com determinada doença e não adquirir imunidade, a possibilidade de contágio se comparado com outros indivíduos da rede é a mesma, então ele permite múltiplos contágios.

No modelo SIS, as doenças não conferem imunidade para as vítimas após a recuperação da infecção, ou conferem somente imunidade limitada, então a infecção pode ser causada mais de uma vez.

No modelo SIRS após o indivíduo recuperar da infecção ele ganha imunidade, como no modelo SIR, porém esta imunidade é apenas temporária e depois de um certo período de tempo a perdem tornando-se novamente susceptíveis a novas contaminações.

2.2.2 Independent Cascade Model

O modelo Independent Cascade Model é um modelo que é utilizado quando temos um grafo direcionado G(V, E), onde temos sementes que disseminam algum conteúdo. Cada aresta tem uma probabilidade de ativação dos próximos nós, por exemplo, se o vértice u é ativado, então ele tem uma probabilidade p de ativação do vértice v, pela probabilidade da aresta seguinte, referente à ele para os próximos nós que tem ligação com ele (WANG; CHEN; WANG, 2012).

Em (WANG; CHEN; WANG, 2012), vemos que o independent cascade model é um modelo do qual cada aresta do grafo tem uma probabilidade de propagação pp(u, v), que é a probabilidade de que o nó u ative de forma independente (à partir da influência que ele tem sobre os outros nós) o nó v (v pode ser qualquer um dos nós que estão conectados a u e não estão ativados) no passo t+1, se u tiver sido ativado no passo t. Este modelo foi o escolhido para ser utilizado, pois cada pessoa, ou vértice numa rede social, tem uma determinada influência sobre os nós que estão ligados à ele, gerando assim uma dependência do prosseguimento de propagação do boato na rede por aquele vértice, e em redes reais há pessoas que detêm maior poder de influenciar pessoas em determinados assuntos do que outras, essa semelhança foi o motivo da escolha do modelo.

Em (WU; WANG, 2014) , é descrito que este modelo é o caso mais simples de modelos de propagação em cascata, pois pode ser modelado de forma simples e é baseado na caracteristica do poder de influência que cada vértice tem sobre os outros vértices a ele conectados.

2.3 Redes Sociais

Em (NEWMAN, 2010) é visto que redes sociais são redes nas quais queremos representar pessoas ou organizações e algum tipo de interação entre as mesmas. As pessoas são os vértices e a interação entre elas são as arestas que interligam um indivíduo da rede à outro, também podemos dizer que os vértices são atores da rede e as arestas os laços entre eles.

Normalmente redes sociais remetem a ideia de redes sociais online, tais como as que estamos acostumados a usar diariamente, além disso, redes sociais tem um âmbito que abrange qualquer interação entre pessoas. No entanto, o estudo e análise de dessas redes é algo que foi estudado e tem registros na literatura desde o século 19, onde começou o interesse no estudo da dinamicidade das interações sociais entre grupos de pessoas.

No domínio de redes sociais temos muitas definições possíveis para arestas, e podem ser levantados muitos significados dependendo da rede que estamos estudando, e no que estamos interessados em responder. O significado das arestas podem ser imensos, seja amizade, relações profissionais, transações monetárias e outras.

Hoje a importância das redes sociais é crucial e essa importância só está crescendo, nossa sociedade está cada dia mais dependente da mesma. São caracterizadas primariamente pela autogeração de seu desenho, pela horizontalidade e descentralização com que são formadas após sua criação.

Na sociedade em que vivemos e com a grande demanda de dados que são produzidos na rede atualmente, vemos que isso pode ser fonte diversos estudos, podendo ser

aproveitadas pela ciência, o poder da análise de rede social é imenso, pois temos uma grande quantidade de problemas a serem resolvidos, em comunidades e na população em geral, redes de contatos profissionais e negócios, colaboração de cientistas, contatos de rede sexual, redes biológicas, entre outros.

Alguns pontos que podem ser destacados para a importância das redes sociais são o compartilhamento de informações, interesses e esforços em busca de objetivos que sejam comuns, podendo estes ter diferentes formas, caracterizando assim comunidades dentro das redes sociais, em busca de pessoas similares e de interesses afins, outro ponto importante das redes é a democratização e mobilização social que podem abranger.

2.4 Nomes Canônicos das Cascatas

As redes complexas simples, da qual é usufruída de suas características neste trabalho são modeladas por um conjunto de vértices e arestas, após a aplicação do algoritmo de espalhamento são criadas diversas árvores, das quais não tem ciclos, ou seja, percorrendo-a a partir da raíz, ou primeiro nó representado sempre chegará em arestas folha, que não tem continuidade na rede. Estas árvores podem ter diversas formas, com muitos ou poucos vértices e uma forma de dar nomes a elas é encontrando um nome canônico para representa-las. Porém duas árvores que tem as mesmas ramificações só que para lados opostos são diferentes visualmente, mas correspondem a duas estruturas iguais, ou seja são isomorfas.

Para tratar o isomorfismo de árvores foi necessária a utilização de um algoritmo proposto por (SMAL, 2008), que está explicado na seção 5.3 denominado *Assign Canonical Names* que dá nomes as cascatas formadas e fazendo o tratamento caso árvores sejam isomorfas, para fins de comparação posterior.

Cada nó tem a representação com os valores "10", porém se ele for nó pai, ou seja, caso tenha descendentes diretos ou indiretos, receberá o nome dos filhos concatenados e assim sucessivamente. No exemplo da árvore abaixo o vértice D tem como seu nome "10", e o vértice C que é pai de D contém o nome do filho no interior do seu nome, que no caso é "10", resumindo, o nome do pai será "1 + concatenação dos nomes dos filhos + 0", como podemos ver no vértice A a representação correspondente é: "1 10 1100 0", o segundo e o terceiro caracter são resultados do vértice B e os caracteres "1100"em A são resultados da concatenação do vértice C seguindo assim de forma recursiva até chegar no vértice que iniciou esta rede.

Para efetuar o tratamento caso duas árvores sejam isomorfas, é realizada uma ordenação no nome dos filhos, na imagem abaixo temos um exemplo, no qual na árvore da esquerda o valor de B é menor que o valor de C, no qual A recebe a seguinte concatenação "1 + nome de B + nome de C + 0", no caso da árvore á direita, o primeiro filho é b e

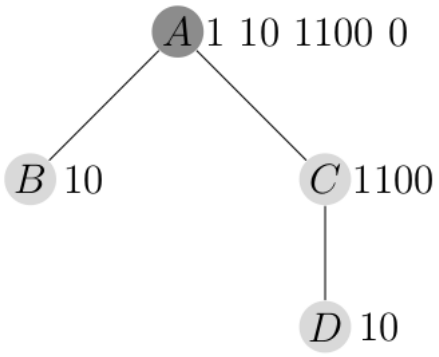

Figura 1 – Nome canônico aplicado a uma árvore com quatro vértices.

o segundo é c, seguindo a lógica, a deveria receber o nome "1 + nome de b + nome de c + 0", só que isto não ocorre pois é realizada uma ordenação no nome dos filhos e consequentemente a recebe "1 + nome de c + nome de b + 0". Verificamos assim que os nomes de A na árvore da esquerda e o nome de a na árvore posicionada à direita são iguais, resolvendo o problema de isomorfismo.

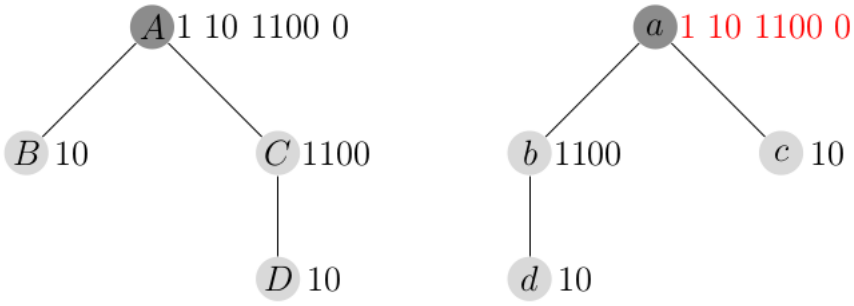

Figura 2 – Nomes canônicos aplicados a duas árvores isomorfas.

3 Trabalhos relacionados

Com o objetivo de estudar os processos de espalhamento de informações *online* e analisá-los para obter algum padrão nas redes sociais, foram estudados uma série de trabalhos que estão relacionados a difusão de informação em redes. Esse estudo visa colaborar para o embasamento e a contextualização do presente trabalho, permitindo uma comparação entre resultados e técnicas aplicadas na área.

Em outros estudos já realizados sobre espalhamento de informações em redes sociais e análise de redes, são sempre levadas em consideração a estrutura que a rede toma após a informação ser semeada, podendo a informação "morrer"prematuramente (ou seja, não ser difundida entre os nós) ou ser propagada entre os nós de alguma forma. Diversos fatores podem ser apontados como causas para a propagação ou não de uma informação, como a importância do nó que semeia ou que compartilha o dado.

Em (DASGUPTA et al., 2008) foi traçado perfil para usuários de uma rede de telefonia, com um valor de função objetivo para cada usuário buscando assim maximizar os lucros da empresa e diminuir fraudes. utilizando da influência de cada nó e uma energia acumulada pelo nó, que terá parte repassada para seus nós filhos. E no final foi concluído que os nós que deixaram a empresa de telefonia não dependem somente das relações com outros indivíduos que migraram de companhia, mas também das relações estruturais presentes entre eles na rede social e que o custo para manter clientes é menor do que o custo para adquirir novos clientes. Notou-se também uma característica de que maioria dos nós da rede conhecem pouca gente, e minoria conhece muita gente, neste caso, a influencia dos nós com muitos conhecidos pode ser maior, por exercer contato com maior quantidade de pessoas.

Em (SCHNEIDER et al., 2013) podemos encontrar outra aplicação real onde podemos verificar padrões à partir da mobilidade urbana, foram feitos mapeamentos dos pontos onde as pessoas frequentavam e vendo a frequência diária com que esses eventos ocorriam, foi possível obter padrões e verificar comportamentos humanos, tais como que todo ser humano tenta minimizar o caminho percorrido, em função de características sócio-econômicas (ZIPF, 1949). Grande maioria das pessoas (cerca de 90%) visita pelo menos sete locais numa base diária, todas iniciando de um local e no final do dia retornando ao mesmo lugar, visto a necessidade de sono e visto que até a mobilidade humana contém padrões, em redes sociais não é diferente, criamos rotinas, compartilhamentos dados, criamos informação e tudo isso possivelmente também tem padrões de formação e espalhamento.

Segundo Anupam Joshi (GOEL; WATTS; GOLDSTEIN, 2012), nos últimos anos,

o aumento da disponibilidade de dados interação social *online* tem oferecido novas oportunidades para mapear a estrutura da rede de processos de difusão. Onde podem ser aproveitados os dados de difusão para verificar qual a adoção da informação na rede, se é semelhante com algo viral ou não. E para identificar características genéricas de estrutura de difusão online, foram realizados sete estudos de diferentes fontes para obter melhores resultados. Neste estudo foi visto que há muitos fatores externos que influenciam no espalhamento de informações, tais como contato pessoal, mudança de meio de comunicação e outros. À partir dos resultados obtidos por (GOEL; WATTS; GOLDSTEIN, 2012), poderíamos utilizar dados de alguma rede social online e capturar algum boato logo de início, e ver o comportamento e a estrutura de difusão que será gerada.

Outro estudo, que pode ser fortemente relacionado ao presente trabalho, foi realizado por Dow *et al.* e explora "*Facebook cascades*" (DOW; ADAMIC; FRIGGERI, 2013), e analisa o comportamento de duas cascatas a partir do compartilhamento de imagens na rede por dois indivíduos bastante distintos. Uma das imagens foi semada por Michelle Obama, esposa de Barack Obama (e essa imagem foi chamada no trabalho de "Obama Victory Picture", ou OVP) e contém uma imagem comemorativa de sua eleição em seu primeiro mandato, em 2008. A outra imagem foi compartilhada por um desconhecido norueguês (e essa imagem foi chamada de "Million Like Meme", ou MLM no trabalho) que fez uma brincadeira com sua amiga, pedindo compartilhamento da imagem. Apesar da ocorrência de cascatas em ambo compartilhamentos, as características observadas são completamente diferentes. Como o compartilhamento de OVP foi feito por uma pessoa notadamente mais influente, a primeira dama dos Estados Unidos da América, houve um alto índice de interação entre os usuários com o post original desde o início e essa interação permaneceu alta durante bastante tempo. Alguns outros usuários tiveram forte influência nos padrões de compartilhamento, o que pôde ser observado, por exemplo, quando a cantora Alicia Keys compartilhou a imagem. Os parâmetros usados para identificar a disseminação foram as interações possíveis pela rede social Facebook, como curtidas, compartilhamentos e comentários. Foram observadas diferenças nos públicos-alvo da imagem e percebeu-se que a OVP atingiu muito o público feminino e pessoas mais velhas, possivelmente pela forte influência de Michelle Obama entre essas pessoas. Por outro lado, a imagem MLM foi compartilhada, majoritariamente, por homens e jovens. Ainda no trabalho de Dow *et al.*, as regiões geográficas atingidas pelas cascatas e o comportamento assumido ao longo do tempo foram analisados, assim como diversos outros fatores que podem influenciar no espalhamento da imagem, tal como a formação de grandes "*hubs*" que compartilham a imagem e a propagam ainda mais, por serem mais propensos a compartilhamentos.

Tomando o exemplo de (DOW; ADAMIC; FRIGGERI, 2013), podemos analisar características de influência das pessoas que espalham algum boato, para verificar qual o alcance que isto pode tomar, calculando o espalhamento dependendo dos nós semente e

dos que compartilham este conteúdo posteriormente, para ver a proporção que o boato toma.

Em (BAKSHY; KARRER; ADAMIC, 2009) foi feito um estudo de como os conteúdos são adotados, o que leva alguém a adotar algum conteúdo, compartilhar com outras pessoas e dissemina-lo. As amizades dos indivíduos exercem grande diferença na hora de aderir algum conteúdo ou produto, se dois indivíduos tem um laço forte, a chance do conteúdo de um ser disseminado pelo outro, é muito alta, ou seja, se os usuários tiverem muita interação, isto é muito provável de acontecer. Para planos de telefonia, uma nova conta é feita pelas necessidades, se maioria dos amigos ou parentes de alguém tiverem alguma operadora, provavelmente esta será a mesma escolhida.

Ainda em (BAKSHY; KARRER; ADAMIC, 2009) percebeu-se que os amigos detêm muita influencia, e quanto mais adeptos aderirem a chance de propagação do conteúdo aumenta. Em alguns casos, a influência fica concentrada em algumas pessoas, como empresas que trabalham com *marketing*, porém a maioria dos indivíduos tem papel insignificante. Foi calculada a entropia dos usuários responsáveis por transferências e comparado com um modelo aleatório gerado, e entropia obtida da distribuição real é menor que a aleatória, provando assim que a rede é mais concentrada do que o esperado se os adotantes anteriores tivessem igual probabilidade. Normalmente as primeiras pessoas que adotam um conteúdo são mais suscetíveis a adotar muitos conteúdos, não exercendo assim muita influência por não selecionar bem o que compartilham. A partir deste trabalho podemos estudar qual a capacidade dos nós de transmitirem alguma informação na rede avaliando seus atributos como número de amigos e fieldade com outros nós da rede, para disseminarem a mesma informação.

Vale ressaltar que há algumas linhas de pesquisa comumente usadas em influência e propagação de conteúdos, que são os trabalhos que investigam o comportamento de cascatas específicas, normalmente sendo estudos de casos, como em (DOW; ADAMIC; FRIGGERI, 2013), trabalhos que exploram a difusão e seu alcance e trabalhos que investigam a formação de cascatas em geral, como feito por (GOEL; WATTS; GOLDSTEIN, 2012). O presente trabalho se baseia nas duas últimas linhas de pesquisa citadas somadas.

4 Metodologia

Neste capítulo é apresentada a metodologia que foi usada para desenvolvimento do presente trabalho de pesquisa. O objetivo principal é simular espalhamento em redes sociais e posteriormente observar quais são os padrões de espalhamento mais identificados à partir de alguns vértices que efetuam o lançamento deste conteúdo. Para explicar a metodologia foram seguidos alguns passos explicados nas seções abaixo:

1. **Revisão da Literatura:** A revisão de literatura foi feita para estudar os algoritmos à serem utilizados, assim como as métricas e técnicas já aplicadas por outros autores.
2. **Seleção de Métricas Utilizadas:** Algumas métricas foram utilizadas à partir do paper (JUNG; HEO; CHEN, 2012).
3. **Desenvolvimento dos Algoritmos:** Desenvolvimento dos algoritmos de cascateamento e extração de padrões dos nós semente.
4. **Seleção das Bases de Dados:** Seleção de algumas bases de dados para testes.
5. **Análise de Resultados:** Análise de resultados à partir dos padrões obtidos e dos espalhamentos gerados sobre as bases de dados selecionadas.

4.1 Revisão da Literatura

Como apresentado no capítulo 3, algumas pesquisas que abordam temas semelhantes e que foram utilizados como base para consolidar o presente trabalho, visto que o objetivo principal é fazer um estudo sobre os padrões que mais aparecem em redes sociais e já existem pesquisas sobre isso na literatura propondo diferentes métricas para utilizar no espalhamento a ser simulado, e algumas delas deviam ser escolhidas para serem utilizadas, assim como o algoritmo para espalhamento e formação das cascatas.

4.2 Seleção de Métricas e dos Algoritmos Utilizados

Para o desenvolvimento da pesquisa era necessário fazer a escolha de alguns algoritmos para serem utilizados, tais como qual algoritmo seria aplicado para fazer o espalhamento do boato nas redes e também como seria feita a extração dos padrões formados pelos espalhamentos. Além disso, foi primordial selecionar métricas para determinar como será feito o espalhamento, pois para um vértice i tentar influenciar um vértice j precisava determinar uma probabilidade deste evento ocorrer e esta probabilidade pode ser calculada de n formas diferentes.

4.2.1 Propagação de Influência

Para o presente trabalho, utilizou-se uma metodologia para efetuar a simulação de cascatas que pode ocorrer em redes reais, das quais acontecem quando algum conteúdo, assunto, ou qualquer outra informação é difundida em algum lugar, seja rede social, conversa informal e/ou qualquer outra forma de comunicação, tal como *e-mail*, boca-a-boca, e outras. A propagação é feita quando alguém tem acesso à algum conteúdo divulgado e faz a adoção do mesmo, prosseguindo assim com a cascata e tentando influenciar novas pessoas que estão ligadas à esta que aderiu o conteúdo de alguma forma.

O desenvolvimento da propagação de influência visa construir uma rede dentre as ligações (ou arestas) que contém nas bases de dados escolhidas para análise na qual é formada por alguns nós que são selecionados para lançar algum produto e tentar influenciar as pessoas que estão ligadas a esta que executa o lançamento da informação na rede, formando assim uma cascata.

Cada vértice tem uma probabilidade p de tentar influenciar os vértices que fazem conexão à ele, e o sucesso irá depender da métrica utilizada para probabilidade de influência.

Há diversos algoritmos propostos na literatura para fim de simular espalhamento em redes, propagando cada um de uma forma diferente, alguns exemplos estão na seção 2.2, e o que utilizamos é o *Independent Cascade Model* descrito na seção 2.2.2.

4.2.1.1 Modelo de Propagação Utilizado - ICM

O algoritmo utilizado denominado *Independent Cascade Model* é amplamente utilizado para pesquisas na área de propagação de influência, pois conforme utilizado consegue fazer uma simulação com bons resultados.

Os documentos utilizados contém a ligação de vértices da rede, ou seja, nele estão as informações de conexões entre quaisquer dois vértices do grafo. Para manipulação dos mesmos foi utilizada a biblioteca *Igraph*, que já possui diversas funções implementadas que foram muito úteis na manipulação de informações no trabalho.

As informações do grafo lidas dos arquivos bases são todas armazenadas nas estruturas pré-definidas do *Igraph*, que faz isso com matrizes de adjacência. No entanto, independente da forma como representamos os dados isto não afetará no resultado final do algoritmo.

Após a definição de como são representados os dados, é feito um esquema de escolhas de quais são os vértices que serão as sementes das cascatas, existindo diversas formas de escolher estes vértices, foi pré-estabelecida somente uma, na qual os vértices que tem maior grau de entrada são as sementes, pois se muitas pessoas se conectam a ele, significa que o mesmo tem uma importância maior na rede que outros vértices dos quais

não são tão interessantes para essa mesma quantidade de pessoas.

Estabelecidos quais são os vértices que irão iniciar o espalhamento na rede, somente são percorridos os adjacentes dos mesmos e são efetuadas tentativas de influenciar estes, seguindo assim a execução do algoritmo recursivamente até não terem mais nós para influenciar pelos que já aderiram ao conteúdo propagado.

A probabilidade de influência no *Independent Cascade Model* é algo que foi abordado por (JUNG; HEO; CHEN, 2012) e foram utilizadas duas métricas por ele abordado, na qual a primeira a probabilidade é igual a

$$\frac{1}{GraudeEntradadoVértice}$$

e a segunda consiste em uma probabilidade aleatória, do qual é sorteado um número entre 1 e 3, posteriormente dependendo do resultado a probabilidade será igual a:

- Se o valor sorteado é 1, p = 0,1
- Se o valor sorteado é 2, p = 0,01
- Se o valor sorteado é 3, p = 0,001

4.2.2 Nomes Canônicos das Cascatas

Após encontradas as cascatas formadas pelo ICM, é necessária uma forma de desenvolver os nomes canônicos das árvores formadas em cada semente, para isso é utilizado o algoritmo *AssignCanonicalNames* que é explicado no capítulo 5.

Para conseguir fazer a extração dos nomes canônicos de uma árvore, é necessário modelar como serão representados os dados e a forma mais fácil encontrada foi criar um dicionário, no qual os índices são os nomes canônicos encontrados e o seu valor é a quantidade de nomes canônicos iguais encontrados nas sementes, ou seja, será usada somente a memória necessária, os nomes canônicos não serão perdidos e poderão ser manipulados posteriormente.

Como os nomes canônicos são representados de forma binária como texto, pode ser feita ordenação dos mesmos, já que é necessária esta operação na função, assim como a operação de concatenação.

Cada vértice das subredes formadas pelo ICM terá uma representação canônica, onde à partir dos nós folha, ou que não tem mais continuidade a partir deles, irão retornar na árvore de recursão até o nó semente para atribuir o nomes a ele, por isso a representação é de suma importância.

No caso das sementes, após cada uma ter seu nome encontrado, os semelhantes serão agrupados para verificar a frequência com que aparecem, para fins estatísticos e de análise de resultados.

4.3 Desenvolvimento dos Algoritmos

Com base nas métricas e nos algoritmos encontrados na literatura, foi consolidada uma forma de obter o resultado almejado para o espalhamento e para a síntese dos padrões gerados por cada vértice que irá lançar o boato na rede e após isso agrupar os semelhantes para utilizar estatística verificando onde e em que cenários são mais frequentes cada tipo de padrão. Os detalhes de implementação estão descritos no capítulo 5.

4.4 Seleção das Bases de Dados

Para aplicar os algoritmos implementados, foram utilizadas bases de dados reais que estão disponibilizadas do site (LESKOVEC; KREVL, 2014), foram utilizadas bases de diversos tipos de redes, para fins de verificação dos resultados obtidos diversificando os tipos de rede e efetuando testes em diversos cenários, das menos às mais populosas e com maior interação entre os integrantes da mesma (ou o contrário). A escolha foi feita em diversos tipos de rede, por isso foram selecionadas bases de site de notícias, de avaliação de consumidores e de redes ponto a ponto.

4.5 Análise de Resultados

A análise de resultados é criada a partir de diversas execuções dos algoritmos, executando eles nas bases de dados propostas e coletando seus resultados para posteriormente compara-los. Com a análise de resultados é necessário responder algumas perguntas, tais como:

- Em quais cenários o espalhamento é maior?
- Quais os padrões mais encontrados geralmente nas redes?
- Com pouca quantidade de sementes a rede consegue se propagar facilmente é necessária uma maior quantidade?

Para responder estas perguntas e visualizar os dados referentes aos testes feitos nos algoritmos, a discussão e amostragem dos dados é feita no capítulo 6.

5 Implementação

A implementação realizada neste trabalho seguiu a metodologia descrita no capítulo 4, seguindo com base nas literaturas referenciadas foi possível obter resultados, que serão descritos no capítulo 6.

Os experimentos de execução dos algoritmos durante a realização deste trabalho foram feitos em um *notebook* Dell, com processador Intel Core i5 @ 1.60GHz x 4 com 5,7 GB de memória RAM e um sistema operacional *Ubuntu* 15.04 com 64 bits.

A linguagem de programação utilizada para efetuar o desenvolvimento dos algoritmos foi a linguagem Python na versão 2.7.9 de 02 de abril de 2015. Algumas bibliotecas utilizadas foram de suma importância, principalmente a *Igraph* que é uma coleção de ferramentas de análise de rede, com ênfase na eficiência, portabilidade e facilidade de uso, sendo que é de código aberto e livre (CSARDI; NEPUSZ, 2006). A visualização dos grafos foi feita utilizando a função *plot* do *Igraph*, quando necessário.

A análise de resultados foi feita utilizando a ferramenta *GNUPlot* para a plotagem dos gráficos, este é um programa de linha de comando que plota gráficos de duas ou três dimensões e outros conjuntos de dados (WILLIAMS; KELLEY; many others, 2010). Os gráficos foram programados por meio de *scripts* na linguagem de programação *Perl* facilitando assim o desenvolvimento e chamadas de execuções no decorrer do trabalho.

5.1 Bases de Dados Utilizadas

Para verificar a efetividade do trabalho aqui apresentado é necessário realizar testes em alguns dados extraídos de redes reais, para simular o funcionamento de fato, portanto foram selecionadas algumas bases de dados, que são descritas na seção 4.4.

1. p2p-Gnutella04.txt

   Esta base contém uma sequência de informações da rede de compartilhamento *Gnutella* que é uma rede ponto a ponto de arquivos, os nós representam as pessoas que compartilham os dados e solicitam na topologia e as arestas são as conexões entre os *hosts Gnutella*.

   No arquivo estão contidos os vértices de origem e destino de cada conexão (aresta), a rede é direcionada, ou seja, se houver uma aresta de um vértice para outro não significa que a conexão seja recíproca. A rede é datada de 4 de agosto de 2002.

   - Quantidade de Nós: 10876

- Quantidade de Arestas: 39994
- Fonte da Base de Dados: https://snap.stanford.edu/data/p2p-Gnutella04.html

2. soc-Epinions1.txt

   Esta base contém as informações de uma rede social de confiança entre usuários, no qual eles realizam uma avaliação geral do site Epinions.com que no final serve para mostrar os comentários mais interessantes para cada usuário, conforme a confiança nos outros utilizadores do site.

   - Quantidade de Nós: 75879
   - Quantidade de Arestas: 508837
   - Fonte da Base de Dados: https://snap.stanford.edu/data/soc-Epinions1.html

3. Slashdot0811.txt

   Slashdot é um site de notícias relacionadas a tecnologia e é muito conhecido pela sua comunidade de usuários específica, no site foi introduzido um recurso para marcar usuários como amigos ou inimigos e a rede contém as ligações de amizade ou inimizade entre os usuários do Slashdot.

   - Quantidade de Nós: 77360
   - Quantidade de Arestas: 905468
   - Fonte da Base de Dados: http://snap.stanford.edu/data/soc-Slashdot0811.html

4. Amazon0302.txt

   Esta rede foi obtida a partir do site da Amazon e baseia-se em clientes que compraram itens no *website*, se algum cliente co-adquiria mais de um item, o grafo conterá uma aresta direcionada de um produto para outro.

   - Quantidade de Nós: 77360
   - Quantidade de Arestas: 905468
   - Fonte da Base de Dados: http://snap.stanford.edu/data/soc-Slashdot0811.html

5.2 Modelo *Independent Cascade Model*

O modelo *Independent Cascade Model* que já foi explicado na seção 2.2.2 é o modelo aqui proposto para fazer a difusão de inluência ou propagação de algum conteúdo na rede, a execução deste algoritmo se inicia com alguns nós sementes ou do inglês *seeds*, tais nós são os iniciais, que lançam o conteúdo na rede e à partir deles são realizadas

tentativas de influenciar os nós adjacentes a eles para também compartilharem do mesmo conteúdo.

O modelo tem algumas regras de funcionamento para ter uma padronização, tais como:

- Teremos um conjunto de nós ativos iniciais *seeds*.
- Um vértice pode tentar influenciar seus adjacentes uma única vez, portanto só haverá uma tentativa do vértice i influenciar o vértice j.
- Haverá uma probabilidade de influência p que irá variar conforme a métrica utilizada.
- Se o vértice i conseguir influenciar o vértice j no tempo t, no tempo $t + 1$, j estará ativo.
- O processo segue até que não haja mais nenhuma ativação possível.

O algoritmo tem um atributo que pode ser variado, que é a aprobabilidade de adoção de influência dos vértices, a definição deste parãmetro é de fundamental importância para o algoritmo, pois se for um valor muito alto a rede toda irá rapidamente ser atingida, ou seja, grande maioria dos vértices podem ficar ativos em poucas iterações, viralizando os conteúdos disseminados na rede muito facilmente e em uma rede real isso é bastante difícil de acontecer, visto que vértices pouco influentes irão influenciar pessoas que normalmente não influenciariam em situações reais.

```
Data: Base de Dados de Rede Social
Result: Nós "Contaminados"
Início;
VetorAtivos = sementes(tempo 0);
while Existir nó ativo (no tempo t-1) do
    Percorre vizinhança;
    P.rand(0,1);
    if (P ≥ P(A,B)) e (B não ativo) then
        B é ativado;
        Empilha B em vetorAtivos(t+1);
    end
    VetorAtivos(t) = VetorAtivos(t+1);
end
```

Figura 3 – Pseudocódigo do modelo *Independent Cascade Model*

.

Outro atributo que foi necessário definir é qual parâmetro iria ser utilizado para selecionar os vértices semente, pois na rede os vértices representam relações reais entre as

conexões das redes, porém não contém informações de quais nós lançaram algum conteúdo no grafo, portanto foi preciso escolher algo para definir os vértices iniciais, no nosso caso foi utilizado o grau de saída, ou seja, os nós que detém maior quantidade de ligações de saída são os primeiros a iniciar o espalhamento e assim sucessivamente.

No tópico a seguir são explicadas as métricas utilizadas.

5.2.1 Métricas aplicadas à Probabilidade de Espalhamento

Duas métricas foram implementadas para definir qual a probabilidade de adoção de conteúdo para os vértices adjacentes a um vértice ativo, tais métricas foram utilizadas em (JUNG; HEO; CHEN, 2012) e foram escolhidas pois as duas tem uma relação com o que acontece na realidade, porém em duas situações diferentes.

Os modelos utilizados são os mesmos utilizados por (JUNG; HEO; CHEN, 2012), que são os modelos de cascata ponderada (*weighted cascade*) e um modelo de trivalência (*trivalency*) dos quais são utilizados como padrões pelo autor.

5.2.1.1 Modelo *Weighted Cascade*

Este modelo atribui uma probabilidade de propagação para cada aresta, esta probabilidade é dada por $\mathrm{P}(u,v) = 1 \,/\, \mathrm{d}v$ onde $\mathrm{d}v$ é o grau de entrada do vértice v.

O modelo *Weighted Cascade* pode ser utilizado para explicar o espalhamento de informações em redes sociais, nos casos onde os receptores de informação adotam uma quantidade semelhante de informação, seja qual for seu grau de entrada, como exemplo podemos citar quando todo mundo lê um número semelhante de *tweets* por um dia no Twitter.

5.2.1.2 Modelo *Trivalency*

Este modelo atribui uma probabilidade aleatória à partir de 3 valores [0.001, 0.01, 0.1] para cada aresta dirigida.

A motivação deste modelo é representar o caso em que há diversos tipos de relações pessoais, no nosso caso são 3 relações, uma para cada valor de probabilidade, portanto a probabilidade de propagação irá depender do tipo de relação que o laço tem, e como não sabemos qual o tipo de relação a aleatoriedade é utilizada para fazer a simulação.

5.3 Atruibuição de Nomes Canônicos

No problema abordado no presente trabalho, temos que cada subárvore formada pelos nós raízes irá formar uma cascata, e para vermos se a cascata de um nó semente

é igual à de outro nó deve-se ter alguma forma de comparar essas estruturas e a melhor forma é encontrar os nomes canônicos de cada raíz e efetuar comparações sobre os mesmos.

Cada grafo pode ser representado pelo seu nome canônico, que nada mais é do que uma sequência de 0's e 1's que representam as ligações entre os vértices, assim como explicado na seção 2.4, algoritmo tal que resolve o problema de isomorfismo, duas árvores são isomorfas caso os nomes canônicos das mesmas são iguais, independente se a forma da cascata for diferente, como exemplo podemos citar se uma delas ramifica para a "esquerda"e outra para a "direita", mas se analisarmos podemos ver que elas são iguais em tamanho, quantidade de vértices e nas ramificações. No algoritmo para encontrar nomes canônicos é feita uma ordenação que faz com que o problema do isomorfismo de árvores seja resolvido, pois ordena os vértices filhos atribuindo ao vértice pai a concatenação dos mesmos, juntamente com o *valor* "10"nos extremos da concatenação, relativos ao vértice pai.

Para este trabalho o algoritmo de Aho, Ullman e Hopcroft (AHU) foi de suma importância, pois encontra os nomes canõnicos em um tempo bom em um tempo O(V) e utiliza a história completa do espectro de grau dos descendentes do vértice como um invariante completo.

A idéia principal dos autores do algoritmo AHU é que ele associa a cada vértice uma tupla que é a descrição da história completa dos vértices descendentes à ele, conseguindo assim atribuir o histórico de vértices dele, porém sem saber quais são os vértices em questão.

```
function AssignCanonicalNames(v);
if (v é semente) then
    Nome de v = "10";
end
else
    while Todos os filhos de v NÃO forem percorridos do
        AssignCanonicalNames(v);
    end
end
Ordenar os nomes dos filhos de v;
Concatenar os nomes de todos os filhos de v em temp;
Dar a v o nome 1temp0;
end function
```

Figura 4 – Pseudocódigo da função para encontrar nomes canônicos

.

6 Resultados

No presente capítulo serão apresentados resultados obtidos quando são utilizados os algoritmos para gerar a simulação de espalhamento de boatos e o algoritmo para extrair o nome canônico das subredes formadas.

Para a execução dos algoritmos foi necessária a utilização de algumas bases de dados, das quais o conteúdo nelas inserido são as conexões estabelecidas entre os vértices, que no caso estão modelando pessoas e as arestas (conexões) que estão modelando as interações entre dois indivíduos presentes na rede.

A escolha das bases de dados foi dada pelo tamanho das mesmas, para verificar se em bases pequenas e grandes o resultado é muito discrepante ou se é semelhante, ainda temos os resultados do (GOEL; WATTS; GOLDSTEIN, 2012) para efeitos de comparação.

As bases de dados utilizadas são as descritas na seção 5.1.

6.1 Resultados do Espalhamento Gerado Pelo *ICM*

Os resultados expostos nesta seção foram obtidos através da aplicação do algoritmo *Independent Cascade Model*, que é explicado na seção 2.2.2 e foi rodado sobre as bases de dados *p2p-Gnutella04*, *Slashdot0811* e *soc-Epinions1* das quais tem seus resultados mostrados nos gráficos abaixo. A finalidade é facilitar a compreensão dos efeitos do espalhamento e do comportamento que o mesmo assume em diversas redes e com métricas diferentes.

Dentre as métricas utilizadas, estão os modelos de propagação Trivalência e *Weighted Cascade*, que são explicados na seção 5.2.1 e em vez de utilizar como parâmetro para escolha de sementes somente os nós que tem os maiores graus de entrada, é utilizada também outra característica, que são os vértices que possuem maior transitividade, gerando assim uma comparação entre os dois, assim como também é feita comparação entre os modelos de probabilidade de influência para fins de maior diversificação e investigação de melhores resultados.

Após realizar a execução do modelo ICM, foram executadas cerca de duzentas vezes o algoritmo para cada base de dados, alterando somente a quantidade de nós semente utilizados e retornando como resultado a quantidade de vértices que foram influenciados por alguma semente ou vértice pertencente à cascata formada por algum nó influenciador.

A formação da cascata se dá com n vértices iniciais, denominados sementes, que serão os formadores de opinião, que irão disseminar o conteúdo, após isso eles tentarão

influenciar seus adjacentes com uma probabilidade p, que poderá ter sucesso ou não e dependerá da métrica de p utilizada. Cada semente formará uma cascata ao influenciar outras pessoas, tais influenciados tentarão propagar o conteúdo adquirido para seus adjacentes e assim sucessivamente, até não ter mais ninguém influenciado e todos que já adquiriram o conteúdo terem tentado propagar para todos os nós que formam seu grau de saída.

6.1.1 Sementes Derivadas de Vértices de Maior Grau de Entrada

Nesta seção serão analisados os gráficos resultantes da execução do algoritmo utilizando como base para escolha dos nós sementes os vértices que contém maior grau de entrada, possuindo assim uma maior quantidade de pessoas para iniciar suas cascatas, ou seja, provavelmente estes vértices irão ter um poder maior de influência.

Na rede *p2p-Gnutella04* as iterações iniciaram com 10 sementes, após esta iteração é feito um incremento de mais 10 nós sementes e assim sucessivamente, até chegar à quantidade máxima de pessoas da rede. Como pode-se perceber visualizando a figura 5 que contém um gráfico, no início tem-se a maior proporção de influência do gráfico, até 4000 sementes a inclinação da curva é bem maior que após este valor, ou seja, começa a estagnar a quantidade de influenciados, visto que sobra uma quantidade menor de pessaos para serem influenciadas e os nós iniciais que tinham maior quantidade de conexões de saída já foram utilizados, minimizando assim a continuidade da cascata.

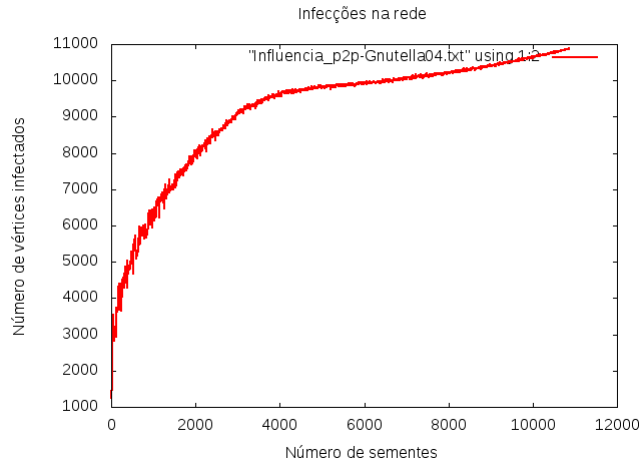

Figura 5 – Modelo ICM aplicado à rede p2p-Gnutella04.

A rede *soc-Epinions1* tem seu resultado aplicado ao *ICM* no gráfico 6, neste gráfico que tem uma dimensão maior, pelo seu tamanho, no qual tem 75879 nós e 508837 arestas no total, sendo assim muito maior que a rede p2p analisada anteriormente. No início do gráfico a influência é maior, bem previsível pela característica dos primeiros nós utilizados para disseminação de persuasão, porém a inclinação da curva em relação ao eixo x diminui em proporção antes da rede p2p, com 5000 vértices ela já declina, e continua declinando,

porém de uma forma muito suave, quase que constante até o final das iterações. O pulo dado na quantidade de sementes entre uma iteração e outra é 250, pois estas redes são muito maiores que a *p2p*.

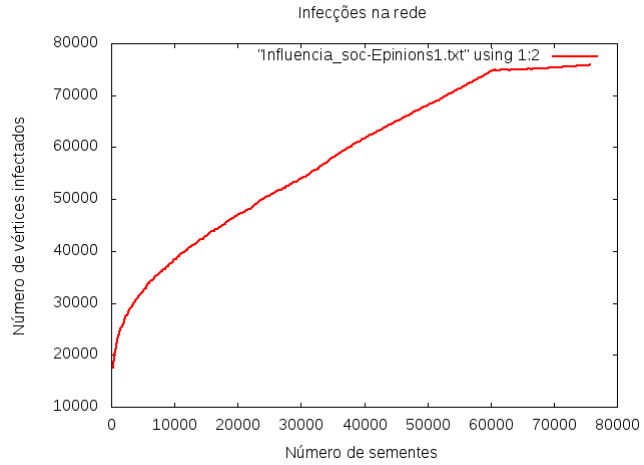

Figura 6 – Modelo ICM aplicado à rede soc-Epinions1.txt

Utilizando a rede *Slashdot0811* verificamos que ela tem comportamento muito semelhante à rede *soc-Epinions1*, porém com uma pequena mudança na curva, que começa a atenuar sua inclinação um pouco antes do que a anterior e segue assim com a curva quase constante até o final das iterações, diferente de *soc-Epinions1* que atenua ainda mais ao final da curva.

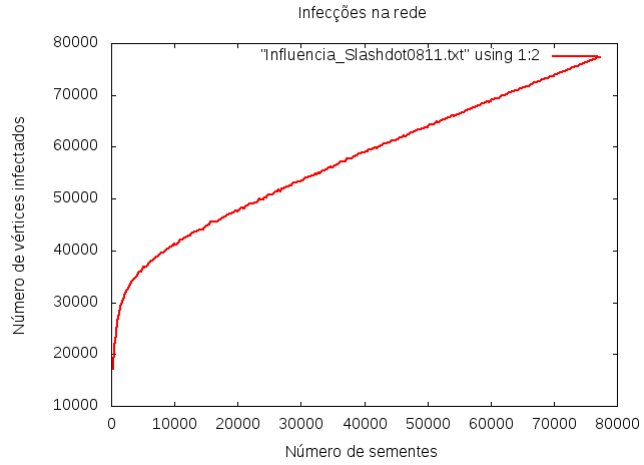

Figura 7 – Modelo ICM aplicado à rede Slashdot0811.txt

6.1.2 Sementes Derivadas de Vértices de Maior Centralidade

A análise feita na presente seção é semelhante à da seção anterior, porém foi feita uma mudança na métrica utilizada para selecionar as sementes disseminadoras de conteúdo, esta seção é baseada na centralidade dos vértices, ou seja, as primeiras sementes

à serem escolhidas são as que tem maior centralidade na rede, indo para as com menor valor e assim sucessivamente, até terminarem as iterações.

Observando o gráfico 5, podemos ver que como a métrica utilizada é baseada na transitividade dos vértices, que não necessariamente o vértice que tem maior transitividade tem maior grau, e consequentemente irá tentar influenciar uma menor quantidade de vizinhos logo de início, gerando assim uma menor probabilidade de gerar cascatas maiores com os primeiros vértices selecionados. A curva segue praticamente como uma reta, variando pouco no decorrer das iterações do início do final, porém decrescendo um pouco a inclinação entre a quantidade de sementes 1500 e 4000, seguindo praticamente como uma reta à partir disso.

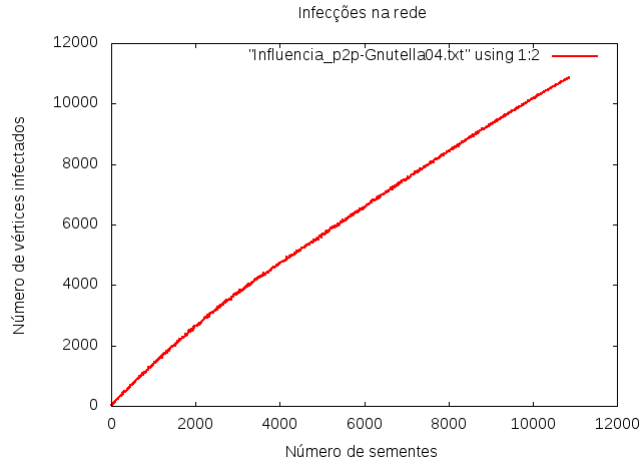

Figura 8 – Modelo ICM com o modelo de probabilidade Trivalência aplicado na rede p2p-Gnutella04.txt

Na base de dados *soc-Epinions1*, segundo o gráfico observamos que o início da curva é dada aproximadamente com 4500 vértices influenciados, isso com 250 vértices sementes iniciais, após isso tem-se um aumento na inclinação da curva em relação ao eixo x aproximando do eixo y, porém isso ocorre mais entre o número de sementes no intervalo de 0 a 10000, como pode ser verificado no gráfico, após isso a curva fica constante, comportando-se semelhante à uma reta.

Na figura abaixo, temos o gráfico referente à base *Slashdot0811* tem o comportamento da curva praticamente igual à *soc-Epinions1*, pois se compararmos os gráficos, do início ao final é verificada uma semelhança muito grande entre ambos, a grande diferença entre os dois gráficos é que no início a rede *Slashdot0811* começa influenciando mais de 10000 nós, enquanto *soc-Epinions1* influencia por volta de 4400. O comportamento a partir do início pode ser observado no decorrer dos gráficos.

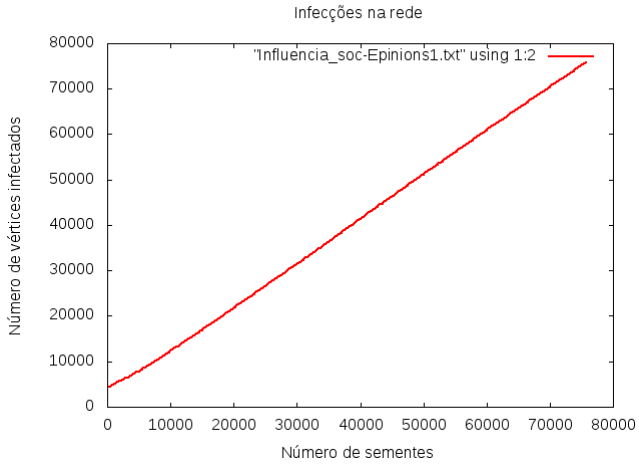

Figura 9 – Modelo ICM com o modelo de probabilidade Trivalência aplicado na rede soc-Epinions1.txt

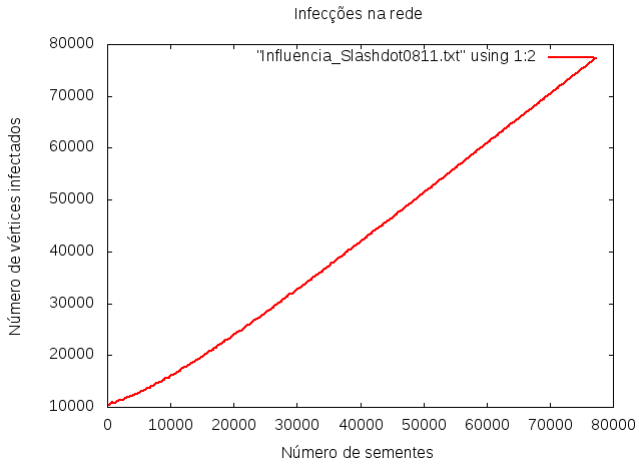

Figura 10 – Modelo ICM com o modelo de probabilidade Trivalência aplicado na rede Slashdot0811.txt

6.2 Padrões Frequentes nas Redes

Nesta seção utilizaremos a Figura 11 como base para verificar quais são os nomes canônicos mais frequentes nas redes para melhores efeitos de visualização. Na imagem são mostrados os padrões juntamente com os nomes canônicos e legendas que serão utilizadas posteriormente nas tabelas.

Os resultados aqui expostos foram obtidos através da aplicação da metodologia descrita no capítulo 4 sobre as bases de dados *p2p-Gnutella04*, *soc-Epinions1*, *Slashdot0811* e *Amazon0302*. As características que analisamos aqui é baseada no trabalho de (GOEL; WATTS; GOLDSTEIN, 2012) que busca encontrar os padrões mais encontrados em redes sociais com aplicação do modelo de propagação de influência *Independent Cascade Model*. É válido ressaltar que neste tipo de experimento a intenção é obter informações úteis e potencialmente novas à respeito de como são formadas as cascatas à partir das pessoas

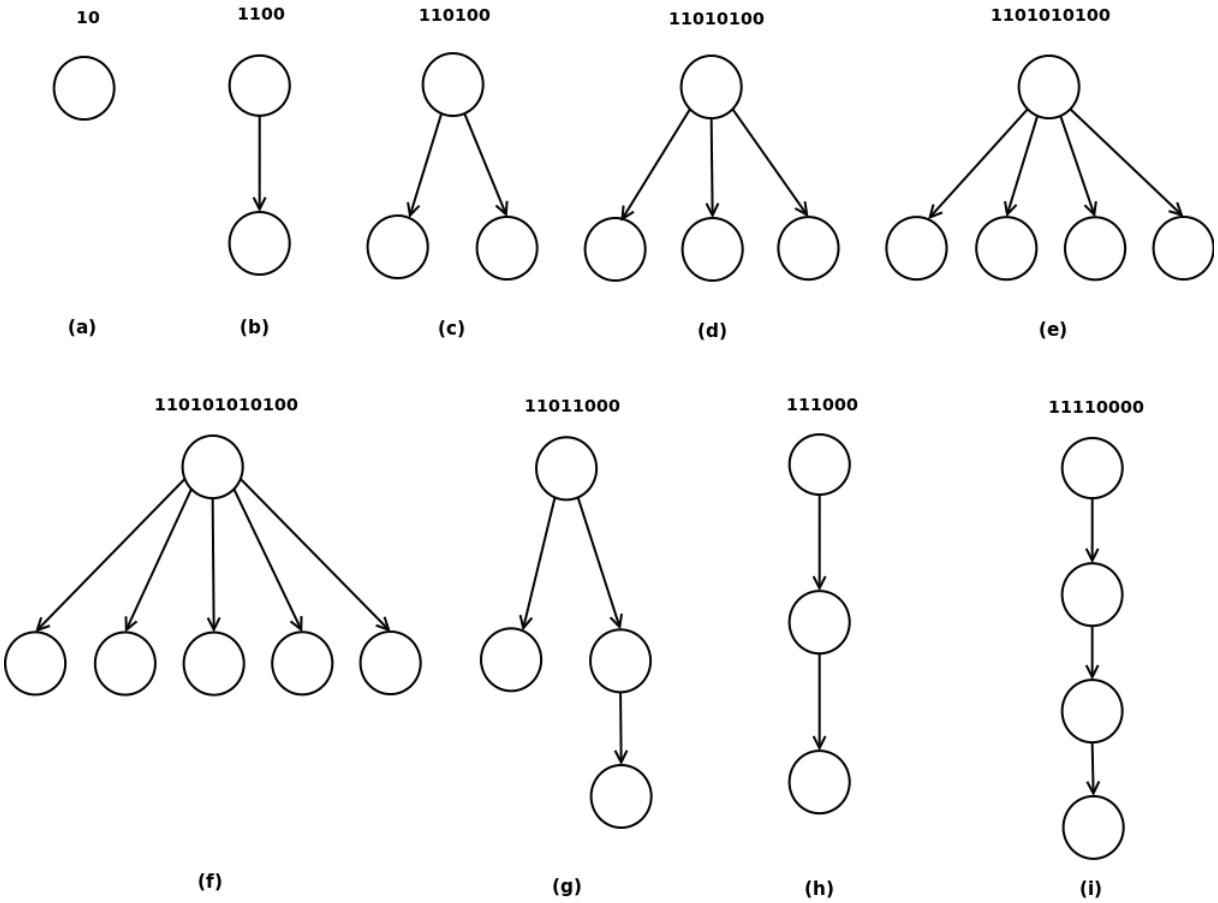

Figura 11 – Padrões mais frequentes com seus respectivos nomes canônicos e legendas

que lançam estes conteúdos em redes reais.

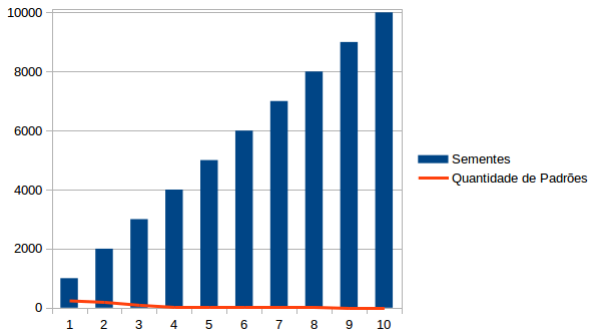

Figura 12 – Gráfico de padrões gerados por número de sementes da base de dados *p2p-Gnutella04*.

Observando a tabela 1 e o gráfico da figura 12 referente a quantidade de nós sementes por quantidade de padrões encontrados na rede *p2p-Gnutella04* podemos observar que quanto maior for o número de *seeders*, menor é a quantia de padrões encontrados, isso se deve ao fato de que, se um vértice lançou o conteúdo, ele não pode ser influenciado sobrando menos pessoas para influenciar.

Na tabela 2 que mostra os padrões mais frequentes, que foi feita com 4 quanti-

Sementes	Padrões
1000	249
2000	198
3000	98
4000	26
5000	12
6000	13
7000	12
8000	11
9000	9
10000	6

Tabela 1 – Tabela de quantidade de sementes e diferentes padrões gerados por *p2p-Gnutella04*.

Tabela 2 – Tabelas de padrões mais frequentes da base de dados *p2p-Gnutella04*

Nº Sementes	Padrão	Frequência
1000	(b)	180
	(c)	149
	(a)	107
	(d)	103
	(e)	45
	(g)	26

Nº Sementes	Padrão	Frequência
4000	(b)	1250
	(a)	1149
	(c)	860
	(d)	416
	(e)	141
	(g)	42

Nº Sementes	Padrão	Frequência
7000	(a)	5046
	(b)	1255
	(c)	464
	(d)	153
	(e)	52
	(f)	14

Nº Sementes	Padrão	Frequência
10000	(a)	9490
	(b)	418
	(c)	65
	(d)	22
	(e)	5
	(f)	1

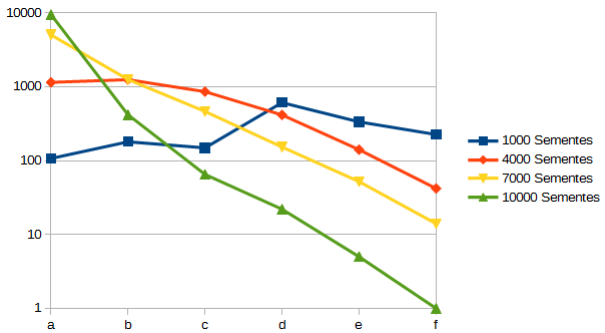

Figura 13 – Padrões encontrados por quantidade de seeders aplicado a rede p2p-Gnutella04

dades diferentes de *seeders*, 1000, 4000, 7000 e 10000, para podermos ter uma noção de como se comporta nesses quatro cenários, vemos que quanto menos nós iniciais ou lançadores de conteúdo, são formadas muitas cascatas diferentes e com maior profundidade e complexidade e conforme vão aumentando as sementes isso vai afunilando, criando menos cascatas e com menor complexidade, ou seja, as mais simples vão aparecendo com uma maior frequência.

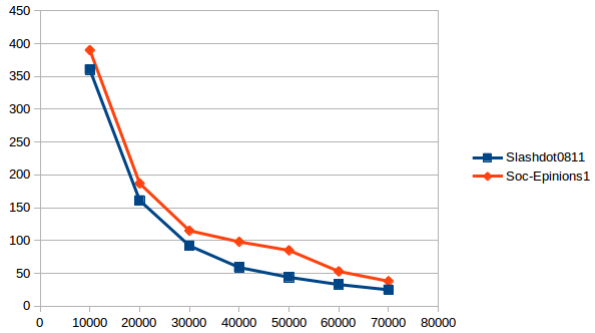

Figura 14 – Gráfico de quantidade de padrões x Número de sementes para as redes da legenda.

Tabela 3 – Quantidade de padrões gerados por sementes para as redes *Slashdot0811* e *soc-Epinions1*

Sementes	**Quantidade de Padrões**	
	Slashdot0811	**soc-Epinions1**
10000	360	390
20000	161	187
30000	92	115
40000	59	98
50000	44	85
60000	33	53
70000	25	38

Na Tabela 3, temos a comparação de duas redes bem parecidas, com uma diferença relativamente pequena de vértices, porém grande de arestas, o grafo de *soc-Epinions1* contém 75,839 vértices e 508,837 arestas e *Slashdot0811* tem 77,360 vértices e 905,468 arestas, esta comparação é válida para mostrar que duas redes parecidas tem quase o mesmo comportamento em termos de padrões e influencia. No gráfico da figura 14 pode-se perceber que em *soc-Epinions1* há mais cascatas diferentes formadas, isso se deve ao fato de existirem mais arestas, gerando uma maior quantidade de opções entre todas as arestas da rede, portanto se um vértice não influenciar o vértice j, outro vértice que tem conexão a ele e poderá efetuar outra tentativa.

Verificando os padrões que ocorrem nestas redes, temos que praticamente todos os padrões mais frequentes são iguais e para cada padrão somente muda a frequência dos padrões, mas neste aspecto também não teve uma discrepância grande.

Tabela 4 – Tabelas de padrões mais frequentes da base de dados *Slashdot0811*

Nº Sementes	Padrão	Frequência
10000	(a)	2928
	(b)	2318
	(c)	1216
	(d)	611
	(e)	335
	(g)	226

Nº Sementes	Padrão	Frequência
30000	(a)	20032
	(b)	5892
	(c)	1733
	(d)	666
	(e)	351
	(f)	203

Nº Sementes	Padrão	Frequência
50000	(a)	43313
	(b)	4281
	(c)	1146
	(d)	457
	(e)	238
	(f)	138

Nº Sementes	Padrão	Frequência
70000	(a)	68244
	(b)	1128
	(c)	282
	(d)	124
	(e)	66
	(f)	51

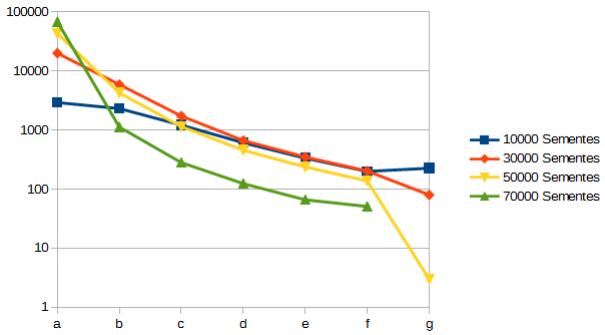

Figura 15 – Padrões encontrados por quantidade de seeders aplicado a rede Slashdot0811

Tabela 5 – Tabelas de padrões mais frequentes da base de dados *soc-Epinions1*

Nº Sementes	Padrão	Frequência
10000	(a)	3841
	(b)	2075
	(c)	1071
	(d)	583
	(e)	342
	(g)	218

Nº Sementes	Padrão	Frequência
30000	(a)	19975
	(b)	6203
	(c)	1846
	(d)	689
	(e)	348
	(f)	186

Nº Sementes	Padrão	Frequência
50000	(a)	42311
	(b)	5326
	(c)	1232
	(d)	484
	(e)	245
	(f)	144

Nº Sementes	Padrão	Frequência
70000	(a)	67936
	(b)	1304
	(c)	313
	(d)	150
	(e)	80
	(f)	40

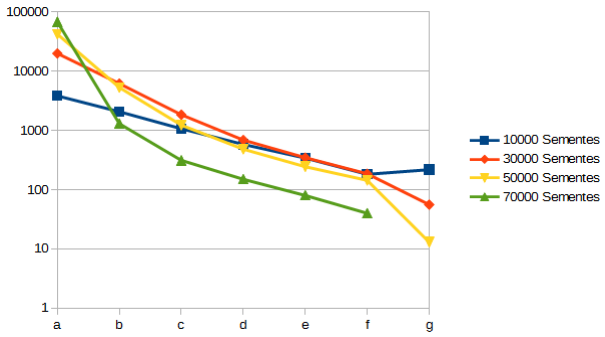

Figura 16 – Padrões encontrados por quantidade de seeders aplicado a rede soc-Epinions1

Na rede Amazon0302 como vemos na tabela 6 e graficamente no gráfico 17 consta uma alternância que no início somente 46 padrões diferentes foram encontrados, até 60,000 a quantidade foi aumentando e posteriormente diminuindo, o motivo para este comportamento provavelmente se deve a quantidade de *seeders* pois com poucos a quantidade é baixa, aumentando os mesmos, a quantidade aumenta, e posteriormente diminui, seguindo este padrão de espalhamento nesta rede.

Tabela 6 – Tabela de padrões gerados por sementes para a rede Amazon0302

Sementes	**Quantidade de Padrões**
1000	46
30000	606
60000	751
90000	688
120000	575
150000	454
180000	329
210000	186
240000	32

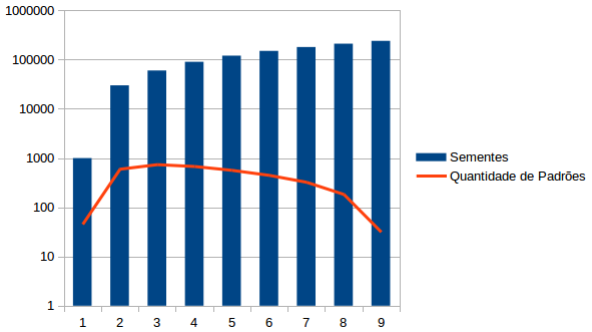

Figura 17 – Quantidade de padrões gerados por quantidade de sementes na rede Amazon0302 (Gráfico na escala log).

Os padrões mais recorrentes, são os mais simples, como vemos na tabela 7, os mais encontrados são frequentes em todas as redes, que são os padrões, "10", "1100"e "111000",

respectivamente, que são os mais encontrados, somente com entrada de 240,000 sementes que o terceiro padrão mais encontrado não é "111000"e sim "110100". Mas no geral, a distribuição de padrões nas redes é muito parecida.

Tabela 7 – Tabelas de padrões mais frequentes da base de dados *Amazon0302*

Nº Sementes	**Padrão**	**Frequência**
10000	(a)	826
	(b)	65
	(h)	27
	(c)	15
	(i)	7
	(j)	5

Nº Sementes	**Padrão**	**Frequência**
60000	(a)	44010
	(b)	8269
	(h)	2119
	(c)	1107
	(i)	593
	(g)	580

Nº Sementes	**Padrão**	**Frequência**
150000	(a)	125495
	(b)	15895
	(h)	2867
	(c)	1788
	(g)	699
	(i)	986

Nº Sementes	**Padrão**	**Frequência**
240000	(a)	224587
	(b)	12807
	(c)	1143
	(h)	855
	(g)	194
	(d)	119

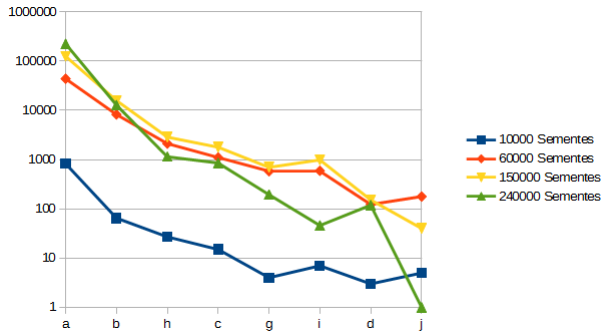

Figura 18 – Padrões encontrados por quantidade de seeders aplicado a rede Amazon0302

7 Conclusão

O presente trabalho teve como principal objetivo a análise dos espalhamentos gerados pelo modelo *Independent Cascade Model* e a partir das cascatas obtidas por cada semente após o espalhamento ser realizado, cada cascata obteve um nome canônico referente à si mesma. Para isso, foram aplicadas duas metodologias de probabilidade de espalhamento, caracterizando diferentes cascatas a partir disso.

A abordagem de espalhamento utiliza o algoritmo ICM que é muito utilizado para propagação de influência. As probabilidades podem ser modeladas de diferentes formas, no caso deste trabalho foram utilizadas dois modelos, o modelo *trivalency* e o *weighted cascade*, um utiliza aleatoriedade e o outro o grau de entrada do vértice.

A identificação dos nomes canônicos das cascatas é feita com o algoritmo *Assign Canonical Names* encontrado em 5.3 com o intuito de agrupar os semelhantes e achar os padrões que mais são formados em diferentes redes sociais pós simulação de espalhamento feita com o ICM.

A aplicação das metodologias descritas no capítulo 4, resultou em cascatas muito parecidas, mostrando que no geral a forma como o espalhamento é feito é muito parecida e depende somente da forma com que a rede original oriunda de (LESKOVEC; KREVL, 2014) foi formada.

Os padrões encontrados em todas as redes são muito parecidos, pelo menos os mais frequentes em todas elas e em todas as situações foram praticamente iguais, mudando em poucas situações, mas isso pode nos mostrar que no geral se pegarmos os usuários que tem maior quantidade de conexões à ele são mais interessantes para tentar influenciar os nós adjacentes do que outros possíveis nós *seeders*, visto que se os mesmos tem maior quantidade de conexões incidindo neles, é porque existe um interesse mútuo dos nós que os conectam. E temos que essas cascatas normalmente são maiores que as formadas por nós com menores quantidades de conexões de entrada. Entretanto, a maior frequência de padrões são os mais simples, no qual um usuário influencia um ou dois novos usuários a aderirem ao conteúdo postado, sem ter uma sequência grande na cascata, sem grande profundidade ou largura.

Como foram utilizadas 4 tipos de redes, sendo destas 3 distintas e duas delas semelhantes, verifica-se que a aplicação em diferentes cenários pode ser bem parecida, variando um pouco os resultados, dependendo de sua formação e conectividade originais, mas no geral os resultados são muito parecidos e suas análises podem ser bem promissoras se aplicadas a outras bases de dados como de redes sociais como o *twitter*, *facebook* e outras.

Como continuidade deste trabalho, propõe-se:

- Implementar novas formas de espalhamento como os modelos de propagação de doenças, para simular espalhamento de doenças e do modelo *Linear Threshold Model* para difusão de influência.
- Implementar novas formas de probabilidade de influenciar novos nós, assim como foram utilizados os modelos *Trivalency* e *Weighted Cascade*, novos modelos podem ser aplicados e comparados.
- Avaliar a quanlidade dos nós, identificando em que cenários quais tipos de nós exercem maior influência sobre seus adjacentes.
- Inspecionar quais qualidades dos nós sementes podem formar maiores cascatas, com maior quantidade de nós influenciados e quais nós não sementes foram influenciados e deram maior continuidade as cascatas.

Referências

ANDERSON, T. K. et al. Ranking viruses: measures of positional importance within networks define core viruses for rational polyvalent vaccine development. *Bioinformatics*, v. 28, n. 12, p. 1624–1632, 2012. Disponível em: <http://dx.doi.org/10-.1093/bioinformatics/bts181>. Citado na página 12.

BAKSHY, E.; KARRER, B.; ADAMIC, L. A. Social influence and the diffusion of user-created content. In: *Proceedings of the 10th ACM Conference on Electronic Commerce*. New York, NY, USA: ACM, 2009. (EC '09), p. 325–334. ISBN 978-1-60558-458-4. Disponível em: <http://doi.acm.org/10.1145/1566374.1566421>. Citado na página 24.

CSARDI, G.; NEPUSZ, T. The igraph software package for complex network research. *InterJournal*, Complex Systems, p. 1695, 2006. Disponível em: <http://igraph.sf.net>. Citado na página 29.

DASGUPTA, K. et al. Social ties and their relevance to churn in mobile telecom networks. In: *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology*. New York, NY, USA: ACM, 2008. (EDBT '08), p. 668–677. ISBN 978-1-59593-926-5. Disponível em: <http://doi.acm.org/10.1145-/1353343.1353424>. Citado na página 22.

DODDS, P. S.; WATTS, D. J. A generalized model of social and biological contagion. *Journal of Theoretical Biology*, v. 232, n. 4, p. 587–604, fev. 2005. Disponível em: <http://dx.doi.org/10.1016/j.jtbi.2004.09.006>. Citado na página 17.

DOW, P. A.; ADAMIC, L. A.; FRIGGERI, A. The anatomy of large facebook cascades. In: KICIMAN, E. et al. (Ed.). *ICWSM*. The AAAI Press, 2013. ISBN 978-1-57735-610-3. Disponível em: <http://dblp.uni-trier.de/db/conf/icwsm/icwsm2013.htmlDowAF13>. Citado 3 vezes nas páginas 13, 23 e 24.

GHOSHAL, G. *Structural and Dynamical Properties of Complex Networks*. BiblioBazaar, 2011. ISBN 9781243700728. Disponível em: <https://books.google.com-.br/books?id=MdniygAACAAJ>. Citado na página 16.

GOEL, S.; WATTS, D. J.; GOLDSTEIN, D. G. The structure of online diffusion networks. In: *Proceedings of the 13th ACM Conference on Electronic Commerce*. New York, NY, USA: ACM, 2012. (EC '12), p. 623–638. ISBN 978-1-4503-1415-2. Disponível em: <http://doi.acm.org/10.1145/2229012.2229058>. Citado 6 vezes nas páginas 13, 22, 23, 24, 34 e 38.

GRANOVETTER, M. Threshold models of collective behavior. *The American Journal of Sociology*, v. 83, n. 6, p. 1420–1443, 1978. Citado na página 16.

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>. Citado 3 vezes nas páginas 25, 27 e 32.

LESKOVEC, J.; KREVL, A. *SNAP Datasets: Stanford Large Network Dataset Collection*. 2014. http://snap.stanford.edu/data. Citado 2 vezes nas páginas 28 e 45.

NEWMAN, M. *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010. ISBN 0199206651, 9780199206650. Citado 3 vezes nas páginas 15, 17 e 19.

SCHMITH, J. et al. Damage, connectivity and essentiality in protein–protein interaction networks. *Physica A: Statistical Mechanics and its Applications*, v. 349, n. 3, p. 675–684, 2005. Disponível em: <http://EconPapers.repec.org-/RePEc:eee:phsmap:v:349:y:2005:i:3:p:675-684>. Citado na página 12.

SCHNEIDER, C. M. et al. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, The Royal Society, v. 10, n. 84, 2013. ISSN 1742-5689. Citado 2 vezes nas páginas 15 e 22.

SMAL, A. Explanation for 'tree isomorphism' talk. 2008. Disponível em: <http://logic-.pdmi.ras.ru/~smal/files/smal_jass08.pdf>. Citado na página 20.

WANG, C.; CHEN, W.; WANG, Y. Scalable influence maximization for independent cascade model in large-scale social networks. *Data Mining and Knowledge Discovery*, 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>. Citado 3 vezes nas páginas 13, 18 e 19.

WILLIAMS, T.; KELLEY, C.; many others. *Gnuplot 4.4: an interactive plotting program*. 2010. http://gnuplot.sourceforge.net/. Citado na página 29.

WU, J.; WANG, Y. *Opportunistic Mobile Social Networks*. 1st. ed. Boca Raton, FL, USA: CRC Press, Inc., 2014. ISBN 1466594942, 9781466594944. Citado 3 vezes nas páginas 13, 16 e 19.

ZIPF, G. *Human behavior and the principle of least effort: an introduction to human ecology*. Addison-Wesley Press, 1949. Disponível em: <https://books.google.com.br-/books?id=1tx9AAAAIAAJ>. Citado na página 22.