UNIVERSIDADE FEDERAL DE SÃO JOÃO DEL-REI

Gilma Aparecida Santos Campos

Estudo do impacto da seleção de sementes baseada em centralidade e em informações de comunidades sobrepostas em redes complexas

São João del-Rei

2021

UNIVERSIDADE FEDERAL DE SÃO JOÃO DEL-REI

Gilma Aparecida Santos Campos

Estudo do impacto da seleção de sementes baseada em centralidade e em informações de comunidades sobrepostas em redes complexas

Dissertação apresentada como requisito para obtenção do título de mestre em Ciências no Curso de Mestrado do Programa de Pós Graduação em Ciência da Computação da UFSJ.

Orientador: Carolina Ribeiro Xavier

Universidade Federal de São João del-Rei – UFSJ

Mestrado em Ciência da Computação

São João del-Rei

2021

Ficha catalográfica elaborada pela Divisão de Biblioteca (DIBIB)
e Núcleo de Tecnologia da Informação (NTINF) da UFSJ,
com os dados fornecidos pelo(a) autor(a)

C3e

Campos, Gilma A..
Estudo do impacto da seleção de sementes baseada em centralidade e em informações de comunidades sobrepostas em redes complexas / Gilma A. Campos ; orientadora Carolina R. Xavier. -- São João del-Rei, 2020.
95 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciência da Computação) -- Universidade Federal de São João del-Rei, 2020.

1. Difusão em redes complexas. 2. Maximização de Influência. 3. Redes Sociais. I. Xavier, Carolina R., orient. II. Título.

Gilma Aparecida Santos Campos

Estudo do impacto da seleção de sementes baseada em centralidade e em informações de comunidades sobrepostas em redes complexas

Dissertação apresentada como requisito para obtenção do título de mestre em Ciências no Curso de Mestrado do Programa de Pós Graduação em Ciência da Computação da UFSJ.

Trabalho aprovado. São João del-Rei, 14 de dezembro de 2020:

Carolina Ribeiro Xavier
Orientador

Carlos Cristiano Hasenclever Borges
Convidado 1

Vinícius da Fonseca Vieira
Convidado 2

São João del-Rei
2020

Agradecimentos

Este trabalho não seria possível sem o apoio de muitas pessoas, e guardarei tudo isso em meu coração com um profundo agradecimento.

Antes de mais nada, gratidão a minha orientadora, Carolina Ribeiro Xavier, que está comigo neste trabalho mesmo com as dificuldades impostas pela vida. Pela sua paciência diante de minhas limitações e, me guiar além delas. Por me proporcionar a descoberta de novos saberes. Pelas gentilezas e auxilio em todo desenvolvimento.

Agradecimento especial ao colega de mestrado, hoje um amigo querido, José Mauro Ribeiro. Por sua valiosa dedicação, comentários, discussões e motivações que me impulsionaram a ir adiante.

Aos membros da minha banca examinadora, agradeço por tão prontamente aceitarem julgar o trabalho.

A Viviane Oliveira, que trouxe luz, paz e equilíbrio para o meu mundo. Gratidão.

Aos amigos que sei, fervorosamente, torceram por mim ao longo desta jornada e outras.

Agradeço pela minha família. A meu pai ausente deste plano, que me deixou aprendizados ímpares. A minha mãe, força para ir adiante. Meus queridos irmãos por me acolherem como sou. Sobrinhos...

Por fim, agradecimentos a outra pequena 'comunidade' com quem tenho laços infinitamente estreitos. Meu esposo, filhos, nora e neto. A quem devo muito do meu crescimento interior.

“Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito. Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes”. (Marthin Luther King)

Resumo

Alguns cenários onde há disseminação de doenças, opinião e ideias, compartilham de características importantes; eles precisam, por exemplo, de indivíduos envolvidos em um contexto social para se espalharem. Redes sociais têm sido amplamente utilizadas para modelar e simular esses processos. Fazendo-se necessário o estudo da sua estrutura para que a escolha destes indivíduos possa garantir a otimização do processo de difusão. O problema de maximização de influência, proposto para redes sociais, está na identificação de um conjunto de *nós* influentes que iniciem o processo de difusão de tal forma que a influência disseminada seja maximizada. Este trabalho se concentra em comparar o alcance da difusão em dois contextos. O primeiro seleciona indivíduos, baseado em medidas de centralidade, enquanto o segundo escolhe os indivíduos utilizando três critérios baseados em comunidades sobrepostas; selecionando indivíduos em cada comunidade de acordo com o tamanho da comunidade, a seleção de indivíduos que participam de várias comunidades e aqueles que estão presentes em mais de uma comunidade e possuam grau elevado. Uma ampla comparação foi realizada sob os modelos de difusão: Independent Cascading Model e Linear Threshold Model. Variando os parâmetros dos modelos, os resultados mostraram que em alguns cenários o uso de comunidades sobrepostas pode trazer melhorias no alcance da difusão.

Palavras-chaves: Redes complexas. Influência. Seleção de sementes. Comunidades sobrepostas.

Abstract

Some scenarios where diseases, opinions, and ideas spread, share significant characteristics; they need, for example, individuals involved in a social context to spread. Social networks have been widely utilized to design and simulate these processes. It is necessary to examine its structure so that the choice of these individuals can guarantee the optimization of the diffusion process. The problem of maximizing influence, proposed for social networks, lies in the identification of a set of influential nodes that start the diffusion process in such a way that maximize the disseminated influence. This work focuses on comparing the scope of diffusion in two contexts. The first chooses individuals based on measures of centrality, while the second select individuals using three criteria based on overlapping communities; selecting them in each community according to the size of the community, the selection of individuals who participate in several communities and those who are present in more than one community and have a high degree. Performed was a broad comparison under the diffusion models: Independent Cascading Model and Linear Threshold Model. By varying the parameters of the models, the results showed that in some scenarios the use of overlapping communities could promote improvements in the reach of diffusion.

Key-words: Complex networks. Influence. Seed selection. Overlapping communities.

Lista de ilustrações

Figura 1 – Exemplo de uma rede social, membros de um clube de karate. 16
Figura 2 – Diagrama das Pontes de Konigsberg. 17
Figura 3 – Exemplo de redes direcionadas e não direcionadas. 19
Figura 4 – Gráfico de distribuição de grau do exemplo de rede não direcionada. . . 22
Figura 5 – A importância dos *nós* A e B de acordo com seu grau. 23
Figura 6 – Exemplo da importância de um nó de acordo com sua proximidade dos outros nós da rede. 24
Figura 7 – Exemplo de Centralidade por Intermediação. 25
Figura 8 – Exemplo de classificação de páginas *Web* de acordo com *PageRank*. . . 26

Figura 9 – Rede de colaboração de cientistas atuando no Instituto de Santa Fé. . . 29
Figura 10 – Esquema de representação de estrutura de comunidade baseado em agrupamento hierárquico gerado pelo método de Girvan e Newman. . . 29
Figura 11 – Exemplo de um pequeno dendrograma de agrupamento hierárquico. . . 30
Figura 12 – Particionamento realizado pelo algoritmo guloso de Newman, exibindo a estrutura de comunidade da rede Karate Clube. 32
Figura 13 – Visão clássica da estrutura de comunidade. 34
Figura 14 – Ilustra o conceito de comunidades sobrepostas. 34

Figura 15 – A-B Jogo de Coordenação. 39
Figura 16 – Processo de Difusão em forma de S. 43
Figura 17 – Dinâmica do processo de difusão no modelo *Threshold*. 45
Figura 18 – Dinâmica do Modelo *Independent Cascade*. 47

Figura 19 – Ilustra os passos realizados na execução dos experimentos definidos para este trabalho. 48
Figura 20 – CDF das Redes Reais. 52
Figura 21 – CDF das Redes Sintéticas para $\mu_u = 0.1$. 54
Figura 22 – CDF das Redes Sintéticas para $\mu_u = 0,2$. 55

Figura 23 – Resultados alcançados para a rede GA-GrQc sob o modelo ICM. 62
Figura 24 – Resultados alcançados para a rede moreno_innovation sob o modelo ICM. 63
Figura 25 – Resultados alcançados para a rede Polbooks sob o modelo ICM. 64
Figura 26 – Resultados alcançados para a rede soc-Epinions1 sob o modelo ICM. . 65
Figura 27 – Resultados alcançados para a rede soc-sign-epinions sob o modelo ICM. 66
Figura 28 – Resultados alcançados para a rede Wiki-Vote sob o modelo ICM. . . . 67

Figura 29 – Resultados alcançados para a rede sintética LFR5 sob o modelo ICM. . 70
Figura 30 – Resultados alcançados para a rede sintética LFR6 sob o modelo ICM. . 70
Figura 31 – Resultados alcançados para a rede GA-GrQc sob o modelo LTM. . . . 73
Figura 32 – Resultados das execuções para a rede HepTh sob o modelo LTM. . . . 74
Figura 33 – Resultados das execuções para a rede Moreno Innovation sob o modelo LTM. 75
Figura 34 – Resultados das execuções para a rede Polbooks para o modelo LTM. . 76
Figura 35 – Resultados das execuções para a rede soc-Eoinions1 sob o modelo LTM. 77
Figura 36 – Resultados das execuções para a rede soc-sign-epinions sob o modelo LTM. 78
Figura 37 – Resultados das execuções para a rede Wiki-Vote para o modelo LTM. . 79
Figura 38 – Resultados alcançados da difusão sob o modelo *Threshold* quando $\mu_u = 0.1$ e o limite é 0.3 para a rede LFR1. 81
Figura 39 – Resultados alcançados da difusão sob o modelo *Threshold* quando $\mu_u = 0.1$ e o limite é 0.3 para a rede LFR6. 82
Figura 40 – Resultados alcançados da difusão sob o modelo *Threshold* quando $\mu_u = 0.1$ e o limite é 0.3 para a rede LFR7. 83
Figura 41 – Resultados alcançados da difusão sob o modelo *Threshold* quando $\mu_u = 0.1$ e o limite é 0.3 para a rede LFR8. 84

Lista de tabelas

Tabela 1 – Informações Básicas das Redes Reais 51
Tabela 2 – Configurações para geração de redes sintéticas de acordo com o modelo Lanchicinetti et al. 53
Tabela 3 – Informações Básicas das Redes LFR com $\mu_u = 0.1$ 53
Tabela 4 – Informações Básicas das Redes LFR com $\mu_u = 0.2$ 54

Tabela 5 – Resultados para as redes reais com execução sob o Modelo *Independente Cascade Model*, cujo valor do limite (*threshold=0,1* é o utilizado na biblioteca NDlib. As letras A, B e C antes do nome dos métodos de detecção de comunidades (SLPA e DEMON) se referem aos critérios propostos para o trabalho de acordo com as execuções para cada método. Onde se lê BC, é *Betweenness Centrality*; CC, *Closeness Centrality* e DC, *Degree Centrality*. 60
Tabela 6 – Resultados para as redes sintéticas com execução sob o Modelo *Independent Cascade*, cujo valor do limite, (*threshold* = 0.1), o mesmo utilizado na biblioteca NDlib. 68
Tabela 7 – Resultados para as redes reais com execução sob o Modelo *Threshold*, cujo valor do limite (*threshold=0.3* próximo aos 0.25 que a biblioteca NDlib utiliza. 72
Tabela 8 – Resultados para as redes sintéticas com execução sob o Modelo *Threshold*, cujo valor do limite (*threshold=0.3*) próximo aos 0.25 que a biblioteca NDlib utiliza. 80

Tabela 9 – Resultados das execuções de detecção de comunidades para os algoritmos Demon e SLPA das redes de estudo. Para cada rede é mostrado: (n_c), o número de comunidades detectadas pelo algoritmo; (ϵ) parâmetro do método Demon e t e r parâmetros do método SLPA; (Q_{ov}), valor da modularidade para o resultado das partições alcançadas; (OMNI) (*overlapping mutual normalized information*), comparação entre os métodos da qualidade da partição em relação a informação mútua normalizada. 95

Sumário

1 Introdução . . . 13
1.1 Objetivos . . . 14
1.2 Justificativa . . . 15
1.3 Organização do Texto . . . 15

2 Redes, Conceitos e Métricas . . . 16
2.1 Redes . . . 16
2.2 Redes e a Teoria dos Grafos . . . 19
2.3 Medidas Básicas de Redes . . . 20
2.3.1 Grau . . . 20
2.3.2 Distribuição de Grau . . . 21
2.4 Medidas de Centralidade . . . 22
2.4.1 Centralidade de Grau . . . 22
2.4.2 Centralidade de Aproximação . . . 23
2.4.3 Centralidade de Intermediação . . . 24
2.4.4 PageRank . . . 25

3 Detecção de Comunidades . . . 28
3.1 Agrupamento Hierárquico . . . 28
3.1.1 Medida de Qualidade de uma Partição . . . 30
3.2 Detecção de Comunidades com Sobreposição . . . 33
3.2.1 Algoritmos de Detecção de Comunidades baseados no LPA . . . 35
3.2.2 Avaliando Partições de Comunidades com Sobreposição . . . 36

4 Difusão em Redes . . . 38
4.1 Maximização de Influência . . . 39
4.1.1 Trabalhos Relacionados . . . 40
4.2 Modelos de Difusão . . . 42
4.2.1 Modelos *Threshold* . . . 43
4.2.2 Modelos *Independent Cascade* . . . 45

5 Metodologia . . . 48
5.1 Dados: Coleta, Tratamento e Caracterização . . . 49
5.1.1 Coleta de Dados . . . 49
5.1.2 Extração da componente gigante . . . 50
5.1.3 Normalização dos dados . . . 51

5.1.4 Caracterização dos dados 51
5.2 Identificação do conjunto de sementes 54
5.2.1 Identificando sementes de acordo com medidas de centralidade . . . 55
5.2.2 Identificando sementes de acordo com informações de comunidades sobrepostas 55
5.3 Descrição dos Experimentos 57

6 Resultados e Discussões 59
6.1 Resultados da Difusão sob o Modelo ICM 59
6.1.1 Redes Reais 59
6.1.2 Redes Sintéticas 67
6.2 Resultados da Difusão sob o Modelo LTM 71
6.2.1 Redes Reais 71
6.2.2 Redes Sintéticas 79

7 Conclusão e Trabalhos Futuros 85

Referências 87

Anexos 94

ANEXO A Resultados das Execuções dos Algoritmos de Detecção de Comunidades. 95

1 Introdução

O advento da Internet tem desempenhado um papel crucial na revolução das formas como a interação social acontece no mundo. Ocorre que no processo de comunicação, juntamente com compartilhamento de informações e ideias, indivíduos são influenciados em suas crenças, ideias e até mesmo em seu comportamento. A influência social acontece quando a ação de uma pessoa afeta outras. Este conceito é definido por Rashotte (RASHOTTE, 2007) como "mudanças em um indivíduo por meio de sentimentos, atitudes, e comportamento que resultam da interação com outro indivíduo ou grupo". Para além do domínio da Internet, as redes sociais que a habitam tem se tornado um ambiente que permite o estudo de influência social, bem como de processos que maximizam esta influência. Campos de estudos dos mais variados tem utilizado destes processos. A difusão de inovações (marketing) buscando maximizar a adoção de um novo produto (KEMPE, 2005). Estudos da influência em processos eleitorais, onde a decisão de voto para um determinado candidato pode depender de quantos outros já tenham decidido por elegê-lo (GRANOVETTER, 1978). Bem como, estudos de controles para limitar a difusão de vírus ou doenças (epidemiologia) (NEWMAN, 2002).

A literatura apresenta diferentes modelos de difusão de informação, os quais podem ser categorizados dentro de três classes, de acordo com Zhang *et al* (ZHANG, 2014), que são: *threshold models*, *cascading models* e *epidemic models*. Por sua vez, os mesmos ganham algumas variações, mas de maneira geral, a literatura, elege o *Linear Threshold Model* e *Independent Cascade Model* como modelos amplamente utilizados em processos de difusão de informação. Difusão, na maioria das vezes envolve a análise da disseminação de um produto ou ideia em um dado sistema social (ROGERS, 2010). Para tanto, diferentes abordagens em pesquisa de difusão podem ser identificadas na literatura, cada uma delas com foco em aspectos particulares do processo de difusão e adotando diferentes metodologias. Vários domínios de estudo se beneficiam deste campo de pesquisa, tais como; economia, marketing, sociologia e antropologia (JOE, 2010).

O problema de maximização de influência consiste em encontrar um conjunto de potenciais adotantes iniciais (sementes) que ao final de um processo de propagação, um número máximo de influenciados seja atingido. Neste sentido, (GULATI, 1998) apresenta um aspecto importante. Que uma rede pode influenciar a ação de seus membros de duas maneiras; por meio do fluxo e compartilhamento da informação dentro da rede e pelas posições dos mesmos na rede. As alegações do autor trazem a importância do estudo da propriedade chave de muitas redes, sua estrutura de comunidades (GREGORY, 2010), pois estudos nessa área são potencialmente úteis na difusão de informação em redes sociais, Lin *et al* (LIN, 2015). Uma rede com estrutura de comunidade é aquela que pode ser

naturalmente dividida em comunidades, onde dentro destas comunidades as conexões são densas e para outras comunidades são esparsas (LEICHT; NEWMAN, 2008).

Assim, como em Granovetter (GRANOVETTER, 1978) e Bakshy *et al* (BAKSHY, 2012), os autores consideram que ligações entre comunidades funcionam diferentemente daquelas dentro das comunidades, ou seja, indivíduos que compartilham da mesma comunidade possuem laços fortes, mas os laços fracos entre indivíduos de diferentes comunidades são importantes na difusão de novas informações, pois possibilitam maior fluxo de informações entre eles. Dessa forma, no trabalho de (LANCICHINETTI, 2009) é apresentado que comunidade sobreposta é quando um *nó* pode pertencer a mais de um grupo. Intuitivamente pode-se corroborar esta alegação, de que indivíduos que são participantes de mais de uma comunidade tendem a possuir maior acesso a novas informações; e utilizar estes indivíduos como adotantes iniciais pode ser significativo para a maximização de um processo de difusão da informação. Resta assim, buscar formas de encontrar estes indivíduos e avaliar a sua importância em um processo de difusão em redes.

Abordagens descritas na literatura e apoiadas por ferramentas de análise de redes possibilitam detectar estes indivíduos, tais como; métricas de centralidade, entre outras. Por meio destas ferramentas pode-se encontrar usuários influenciadores e estudar como é o fluxo da informação através da rede (JALALI, 2016). Assumindo estes indivíduos como sementes para o processo de difusão deve-se ter como ponto principal a identificação de comunidades disjuntas (XIE, 2013). O autor considera que neste tipo de detecção é assumido que a rede pode ser particionada dentro de regiões (comunidades) densas na qual *nós* tem mais conexões dentro de uma comunidade do que para o resto da rede. No contexto de comunidades, Easley *et al* (EASLEY, 2010) alega que a transmissão de informação/inovação e/ou propagação de um comportamento pode parar quando ele tenta entrar em uma comunidade muito unida dentro da rede. De maneira que, a utilização de *nós* alocados em comunidades sobrepostas pode ser uma boa estratégia para que a difusão não encontre barreiras para adentrar comunidades densamente conectadas. Diante da ampla gama de domínios que necessitam da utilização destes processos, ainda é profícuo estudos que tratem da difusão de informação/inovação.

1.1 Objetivos

O objetivo deste trabalho é comparar processos de difusão em redes para verificar o alcance do fenômeno a partir da divisão de redes em comunidades com sobreposição. Especificamente, pretende-se:

- utilizar métricas de centralidade para selecionar conjuntos de sementes para iniciar o processo de propagação na rede como um todo;

- selecionar conjuntos de sementes a partir de comunidades detectadas por algoritmos com este fim;
- utilizar dois modelos de propagação de influência, *Independente Cascade Model* e *Linear Threshold Model*.

1.2 Justificativa

A essência da Maximização de Influência é identificar sementes de qualidade e recomendá-las aos tomadores de decisão que vão direcionar um processo de propagação para um propósito específico. Mais do que utilizar um modelo de propagação, a estratégia utilizada para selecionar as sementes de adotantes iniciais deve ser altamente consideradas. Como alega Gulati (GULATI, 1998), a posição de indivíduos na rede é um fator a ser considerado para que a propagação aconteça. Apesar dos diversos trabalhos presentes na literatura com este foco, definir sementes para uma difusão ótima ainda é uma atividade difícil e a escolha de sementes baseadas em comunidades com sobreposição não está completamente explorada.

1.3 Organização do Texto

O restante deste trabalho está organizado da seguinte maneira: os Capítulos 2, 3 e 4 apresentam as bases teóricas e empíricas que norteiam o presente trabalho. Mais especificamente, no Capítulo 2 são apresentados conceitos relacionados a redes complexas. O Capítulo 3 apresenta a teoria de estruturas de comunidades em redes e métodos e algoritmos utilizados para particionar redes. O Capítulo 4 descreve fundamentos de propagação de influência, o problema de maximização de influência, o funcionamento dos modelos *Independente Cascade* e *Linear Threshold* e ainda, uma breve revisão de trabalhos relacionados. O Capítulo 5 conta com a metodologia empregada para a execução do estudo. O Capítulo 6 é dedicado a apresentação dos resultados e discussões para o conjunto de redes reais selecionados e os *benchmarks* utilizados são apresentados e discutidos. Por fim, no Capítulo 7 conta com as conclusões do trabalho e propostas de trabalhos futuros.

2 Redes, Conceitos e Métricas

2.1 Redes

A sociedade moderna tem observado, com fascínio, o crescente e complexo conceito de conectividade, ou seja, um padrão que surge de uma rede de itens interconectados (EASLEY, 2010). Redes com este tipo de padrão, são também consideradas sistemas complexos, devido a dificuldade em compreender o comportamento coletivo, bem como conhecer componentes que emergem dos sistemas. Estes sistemas aparecem em diversos tópicos e campos na sociedade (EASLEY, 2010) (BARABASI, 2016) e pesquisadores têm utilizado deles para expandir os estudos nestes tópicos e campos (COHEN; HAVLIN, 2010). O mundo tecnológico propicia o surgimento de tais redes (COHEN; HAVLIN, 2010): redes de comunicação, redes elétricas, redes de linhas aéreas, a internet, etc. A sociedade também está em rede, haja vista, redes de relacionamentos entre indivíduos; relações de trabalho ou *hobbies* em comum. Cidades e países conectados por estradas e redes de linhas aéreas. Propagação de epidemias, carregadas não apenas pelas propriedades dos patógenos, mas também pela estrutura da rede social dentro de uma população (EASLEY, 2010). Assim, vários fenômenos podem ser caracterizados dentro do contexto de redes.

Fundamentalmente uma rede é qualquer coleção de objetos na qual os pares destes objetos são conectados por links (relacionamentos e conexões) (EASLEY, 2010).

A Figura 1, exibe uma rede de indivíduos (Karate Club), clássica na literatura de redes (NEWMAN, 2010) (EASLEY, 2010) estudada pelo antropólogo Wayne Zachary, na década de 70, em uma universidade. A mesma é composta por 34 indivíduos que são exibidas como círculos e linhas que unem pares deles, demonstrando círculos de amizade fora do ambiente do clube.

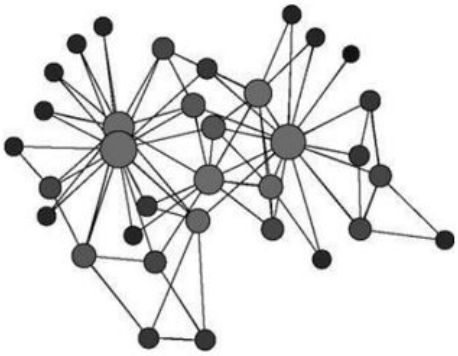


Figura 1: Exemplo de uma rede social, membros de um clube de karate.

Fonte: (NEWMAN, 2010)

Para Barabási (BARABASI, 2016), o crescente interesse em ciência de redes durante a primeira década do século XXI está centrada na descoberta de que apesar da óbvia diversidade de sistemas complexos, a estrutura e a evolução das redes por trás destes são dirigidas por um conjunto comum de leis e princípios fundamentais.

Assim, para descrever redes, matematicamente, é utilizada a teoria dos grafos, que representa as propriedades topológicas essenciais de uma rede, tratando-as como uma coleção de nós e arestas. Por exemplo, em uma rede de computadores, estes podem ser representados por *nós* e, os cabos entre eles pelas arestas. O conceito é bem simples, todavia poderoso, pois à sua simplicidade, ele considera diferentes sistemas complexos, utilizando ferramentas e métodos matemáticos que cabem a todos devido as semelhanças nas propriedades de redes (COHEN; HAVLIN, 2010).

Em (BARABASI, 2016) (COHEN; HAVLIN, 2010) entre outros, é dito que o trabalho de Euler inicia a teoria dos grafos no século XVII, considerado pai do campo da topologia e outros campos matemáticos, com o famoso problema das pontes de Königsberg. O problema é exemplificado na Figura 2, onde pessoas se perguntaram por anos se poderiam percorrer todas as partes da cidade conectadas pelas sete pontes, sem passar por qualquer uma delas duas vezes. Euler compreendeu que o único fator importante nesse problema é a estrutura topológica da rede e, portanto, pode ser simplificada em um problema de grafos, contendo *nós* (partes da cidade) e arestas (pontes).

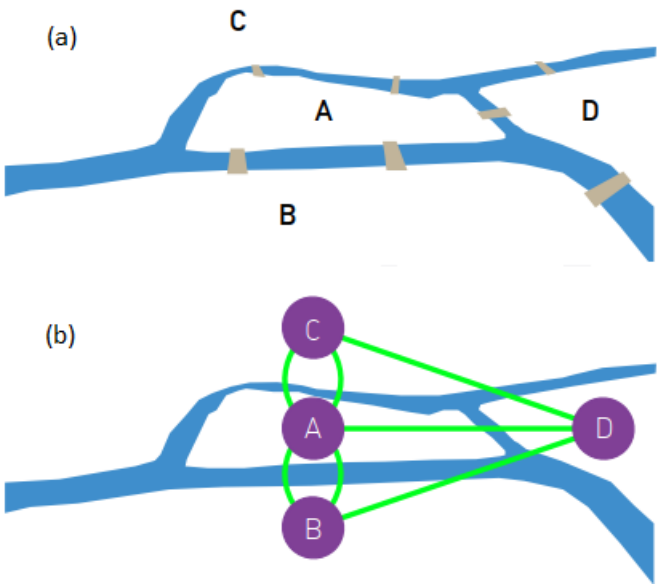


Figura 2: Diagrama das Pontes de Konigsberg.

Fonte: (BARABASI, 2016)

Euler representou cada uma das quatro partes da cidade separadas pelo rio com as letras A, B, C e D, como mostra a Figura 2a. Depois ele conectou com linhas cada uma das partes da cidade que possuía uma ponte entre elas, como na Figura 2b, formando um

grafo. O trabalho provou que o percurso presumido não seria possível, pois se existe um caminho cruzando todas as pontes, nós com número impares de arestas devem ser ponto inicial ou final do caminho. Assim, se um indivíduo chega em um nó com número impar de arestas, ele pode não ter nenhuma aresta não utilizada para deixa-lo, (BARABASI, 2016). Todavia, a teoria de grafos possui natureza estática, enquanto redes, geralmente, são sistemas dinâmicos, pois sofrem mudanças ao longo do tempo (LESKOVEC, 2008) (BARABASI, 2016). Por exemplo, redes tecnológicas mudam diariamente pela adição de novos nós e arestas entre eles, redes biológicas mudam pelos processos evolucionários e ambientais (COHEN; HAVLIN, 2010).

Paul Erdös e Albert Rény, introduziram a teoria de redes aleatórias, que permitiu o tratamento de tais redes, combinando conceitos da teoria dos grafos com ferramentas da teoria da probabilidade (COHEN; HAVLIN, 2010). O modelo define que para um dado número de *nós*, cada par deles tem uma idêntica probabilidade de estar conectado por uma aresta, (ERDŐS; RÉNYI, 1960).

Com o avanço dos computadores, a disponibilidade de dados em grande escala e ferramentas para análise, torna-se conhecido que a teoria de redes aleatórias não descrevem redes do mundo real, (COHEN; HAVLIN, 2010). No trabalho de (BARABASI, 2016), é assumido que a distribuição de arestas de algumas redes não é completamente aleatória, ou seja, as interações entre as proteínas do corpo são governadas por leis bioquímicas (para uma célula funcionar sua arquitetura química não pode ser aleatória). Do mesmo modo, em uma sociedade aleatória, existe a probabilidade de um estudante americano ter entre seus amigos, operários chineses, mais do que seus amigos de classe. Neste sentido, são consideradas algumas correlações menos triviais encontradas nas redes do mundo real (COHEN; HAVLIN, 2010). Para este modelo, a rede cresce pela inclusão de novos *nós* ao longo do tempo. Quando o novo *nó* chega a rede ele se conecta por meio de uma aresta, não aleatoriamente, mas com probabilidade proporcional ao grau do *nó* ao qual ele irá se ligar (LESKOVEC, 2008). Assim, a formação que emerge dessas regras locais faz com que a estrutura da rede seja complexa.

Para além da teoria dos grafos, o estudo de redes, tais como sociais, leva a novas leis físicas, que são decorrentes da nova topologia (COHEN; HAVLIN, 2010). Ou seja, a existência de uma organização profunda por trás delas que reflete em sua estrutura, descrevendo sua arquitetura. Entretanto, este campo auxilia na maneira como as redes podem afetar as propriedades de um sistema, sua representação e principais características (BARABASI, 2016).

2.2 Redes e a Teoria dos Grafos

De acordo com a definição matemática, um grafo é um par de conjuntos *(V, E)*, onde V é um conjunto de vértices (nós do grafo), e E é o conjunto de arestas, também definida por conexão entre os vértices (NEWMAN, 2010) (BARABASI, 2016) (LESKOVEC, 2008) (EASLEY, 2010) (COHEN; HAVLIN, 2010). Uma rede pode assumir conexões direcionadas ou não direcionadas, possuindo ainda múltiplas arestas, bem como '*self loops*' (BARABASI, 2016) (NEWMAN, 2010) (LESKOVEC, 2008).

Arestas podem ter um número variado de características que podem ser utilizadas para realizar a análise de uma rede. Podem ser rotuladas, descrevendo sobre o relacionamento entre os *nós*, possuir peso indicando informação numérica sobre o relacionamento e possuir direção, ou seja, direcionada ou não direcionada (GOLBECK, 2013). Uma aresta é dita não direcionada quando indica um relacionamento mútuo, ou seja, se pares $(i, j) \in E \leftrightarrow (j, i) \in E$ (LESKOVEC, 2008). Por outro lado, direcionada indica um relacionamento que um *nó* tem para com outro, mas que podem possuir reciprocidade ou não (GOLBECK, 2013). A Figura 3 adaptação de (NEWMAN, 2010), exemplifica os tipos de grafos acima referenciados. Nos itens a e b da Figura 3 constam um grafo com 6 nós e suas arestas.

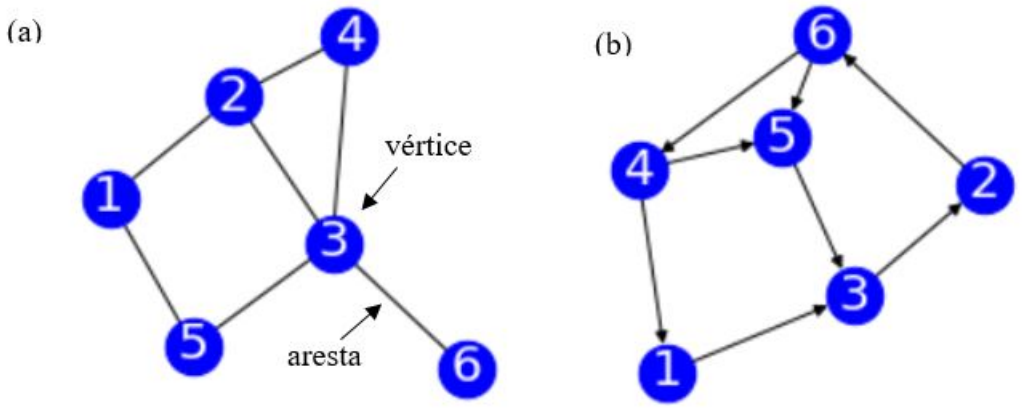


Figura 3: Exemplo de redes direcionadas e não direcionadas.

Fonte: Adaptado de (NEWMAN, 2010)

Para descrever uma rede, é necessário que a organização de suas arestas seja conhecida (BARABASI, 2016). Este objetivo é alcançado com a representação utilizando listas de arestas (*edge list*) (ESTRADA, 2012) (NEWMAN, 2010) (BARABASI, 2016), por exemplo, para a Figura 3a; $(1, 2), (1, 5), (2, 3), (2, 4), (3, 4), (3, 5) e (3, 6)$ uma das mais comuns representações de uma rede (GOLBECK, 2013). Outra representação também muito utilizada é a matriz de adjacência. Assim, para uma rede simples A, com elementos

$A_{i,j}$ tal que (BARABASI, 2016) (NEWMAN, 2010) (ESTRADA, 2012):

$$A_{i,j} = \begin{cases} 1 & \text{se existe uma aresta entre nós i e j} \\ 0 & \text{caso contrário} \end{cases}$$

Em uma matriz de adjacência, os nós são listados em ambas as dimensões da matriz, x e y. Assim, para cada célula na matriz, são alocados valores para indicar se existe ou não uma aresta entre cada par de *nó* (BARABASI, 2016) (NEWMAN, 2010) (GOLBECK, 2013). Para as redes da Figura 3, as matrizes de adjacência seriam dadas como:

$$A = \begin{Bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{Bmatrix}$$

$$A = \begin{Bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{Bmatrix}$$

Da base matemática (teoria dos grafos), cientistas sociais, cientistas da computação e da física vêm colaborando para estabelecer teorias que subsidiem a criação de algoritmos para o cálculo de medidas de rede. Medidas de rede (medidas quantitativas) permitem analise sistemática para que se possa "comparar redes, rastrear mudanças ao longo do tempo e, determinar a posição relativa de indivíduos e grupos dentro da rede" (HANSEN, 2010).

2.3 Medidas Básicas de Redes

2.3.1 Grau

O grau de um *nó* define o número de arestas conectadas a ele, (BARABASI, 2016) (NEWMAN, 2010) (NEWMAN, 2010) (COHEN; HAVLIN, 2010). Para (BARABASI, 2016), por exemplo, em uma rede de telefonia, o grau pode representar o número de indivíduos com o qual um indivíduo em particular tem se comunicado, ou ainda, em uma rede de citações, o número de citações que uma pesquisa recebe em tal rede. Assim, Newman (NEWMAN, 2010) coloca que para uma rede não direcionada com n *nós*, o grau de um *nó* pode ser descrito em função da matriz de adjacência na equação abaixo:

$$K_i = \sum_{i=1}^{n} A_{i,j}.$$

Ou como a soma do grau de todos os nós. Visto que em redes não direcionadas, arestas m são computadas duas vezes $2m$, como o grau de cada *nó* em sua extremidades:

$$2m = \sum_{i=1}^{n} K_i.$$

Também é considerada uma importante propriedade de uma rede, o grau médio (BARABASI, 2016), que é dado pela média aritmética do grau c de todos os *nós*, como expressada na seguinte equação de Newman (NEWMAN, 2010) :

$$c = \frac{2m}{n}$$

A diferença em computar graus dos nós em redes não direcionadas e direcionadas é devido ao fato de redes direcionadas possuírem dois graus, de entrada e de saída, *in-degree* e *out-degree*, respectivamente (NEWMAN, 2010). *In-degree*, k_i^{in}, define o número de arestas que apontam para o nó i, e *out-degree*, k_i^{out}, define o número de arestas apontando do nó i para um outro *nó*. Neste caso, o grau de um *nó* é definido pela equação abaixo (BARABASI, 2016):

$$K_i = K_i^{in} + K_i^{out}$$

Considerando a representação de uma rede por sua matriz de adjacência, os graus, *in-degree* e *out-degree* podem ser descritos da seguinte forma em (NEWMAN, 2010):

$$m = \sum_{i=1}^{n} K_i^{in} = \sum_{j=1}^{n} K_j^{out} = \sum_{i_j} A_{i_j}.$$

O grau médio para uma rede direcionada em (BARABASI, 2016) é dado por:

$$\left(k^{in}\right) = \frac{1}{N}\sum_{i=1}^{N} K_i^{in} = \left(k_i^{out}\right)\frac{1}{N}\sum_{i=1}^{N} K_i^{out} = \frac{L}{N}$$

2.3.2 Distribuição de Grau

Distribuição de grau é uma medida gerada pelo conhecimento do grau de um *nó* (NEWMAN, 2010) (BARABASI, 2016) (COHEN; HAVLIN, 2010). Esta medida fornece a distribuição de frequência dos graus dos *nós*, ou seja, pk define a fração de nós tendo grau k. De forma mais simplista, (GOLBECK, 2013) coloca que, para cada valor de grau, quantos nós o possuem. Redes do mundo real, geralmente, apresentam grau de

distribuição seguindo a distribuição *power-law* (lei de potência), cauda longa, (COHEN; HAVLIN, 2010) (BARABASI, 2016) (NEWMAN, 2010). Seguindo esta distribuição, uma rede possui muitos nós com poucas arestas e alguns poucos com muitas arestas. Na Figura 4 é exibido o grau de distribuição referente a Figura 3a, onde há três *nós* (1,4 e 5) com $k = 2$, um *nó* (2) com $k = 3$, um *nó* (3) com $k = 4$ e o *nó* (6) com $k = 1$.

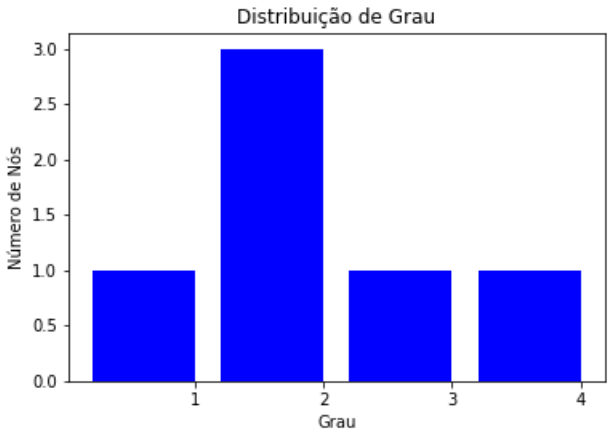


Figura 4: Gráfico de distribuição de grau do exemplo de rede não direcionada.

Fonte: Dados da pesquisa.

2.4 Medidas de Centralidade

Medidas de centralidade são utilizadas para identificar propriedades excepcionais de *nós* em uma rede, ou seja, aqueles que são mais importantes (HANSEN, 2010) (WASSERMAN, 1994)de acordo com um desejado contexto. Do ponto de vista de Hansen (HANSEN, 2010), aqueles que possuem maior poder de influência, os mais eficientes, ou mesmo aqueles que não podem ser substituídos. Nos trabalhos de ambos os autores, a base para estas medidas está na "localização" de um *nó* na rede. Podem ser calculadas, por exemplo, a partir da vizinhança do *nó* (HANSEN, 2010). Wasserman (WASSERMAN, 1994) considera que é um dos primeiros temas, da teoria de grafos, utilizados na análise de redes sociais.

2.4.1 Centralidade de Grau

O grau de conectividade de um *nó*, definido anteriormente, na seção 2.3.1, é provavelmente considerado a mais simples e direta medida de centralidade (WASSERMAN, 1994) (NEWMAN, 2018) (MARTIN; NIEMEYER, 2019) (ESTRADA, 2012). Por exemplo, a partir de uma rede social, é razoável supor que uma pessoa que possui muitos amigos ou conhecidos tem mais influência, mais informação ou mais prestígio que outros com menor grau (NEWMAN, 2018). (ESTRADA, 2012) fornece um outro exemplo. Considerando o efeito de nós centrais como portadores de infecção em uma rede, isto quer

dizer que um link/aresta conectado a eles implicará em infecção. Observando a Figura 5, (AL-TAIE; KADRY, 2017) pode-se verificar que o *nó* B é o que possui maior grau, com oito conexões, seguido do *nó* A, este com 4 conexões. De maneira que, de acordo com o grau o *nó* B possui maior importância.

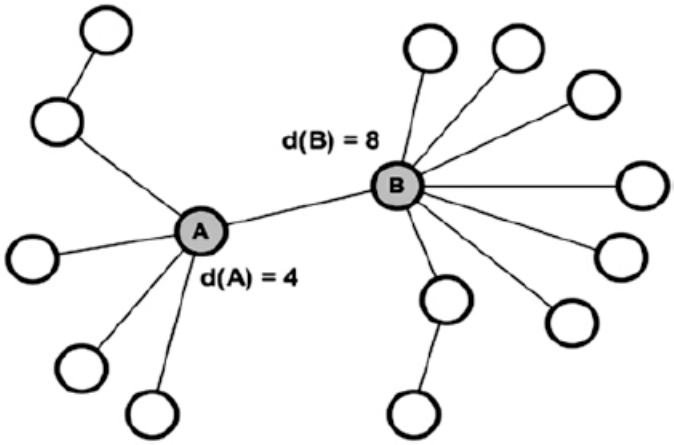


Figura 5: A importância dos *nós* A e B de acordo com seu grau.

Fonte: (AL-TAIE; KADRY, 2017).

2.4.2 Centralidade de Aproximação

A visão da centralidade de aproximação (*closeness*) é baseada na proximidade ou distância, capturando o quão perto um *nó* está de outro na rede e consequentemente, o quão fácil e rápido este *nó* pode alcançar outros (WASSERMAN, 1994) (NEWMAN, 2018). Wasserman (WASSERMAN, 1994) prega que a ideia é de que um *nó* é central se ele pode interagir rapidamente com todos os outros. Assim, esta medida para um dado *nó* é definida como a soma de suas distâncias para todos os outros nós na rede. Isso faz com que seja considerada uma medida de importância para o *nó* quando avalia, por exemplo, quanto tempo e/ou qual o alcance de um processo de difusão deste *nó* para outros, sequencialmente (NEWMAN, 2018), tal como uma cascata de informação (EASLEY, 2010). Na Figura 6, retirada de (BISHT, 2018), o *nó* B é o mais importantes em relação à proximidade, pois ele consegue alcançar o restante dos nós na rede com facilidade.

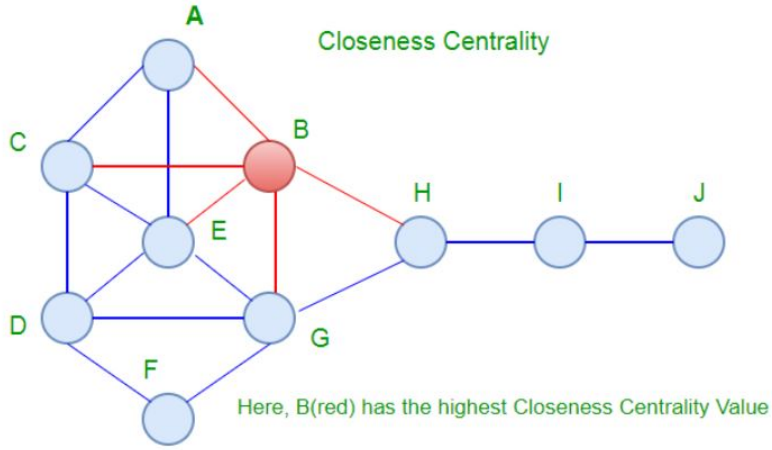


Figura 6: Exemplo da importância de um nó de acordo com sua proximidade dos outros nós da rede.

Fonte: (BISHT, 2018).

2.4.3 Centralidade de Intermediação

Centralidade *betweenness* (intermediação) captura o papel de um *nó* como conector ou ponte entre outros grupos de *nós* (ZAFARANI, 2014) (HANSEN, 2010). Esta medida busca definir o quanto afeta a funcionalidade da rede, um *nó* cujo papel é servir de ponte entre outros *nós* na rede (WASSERMAN, 1994). Ou ainda, ela mede o controle de um *nó* sobre toda informação que flui por entre todos os pares de *nós* em seu componente (NEWMAN, 2018), o que faz destes *nós* importantes usuários, tendo grande influência no que acontece na rede (VARLAMIS, 2010) (GOLBECK, 2013). Na Figura 7, exemplificado em (HANSEN, 2010) se torna evidente que o *nó* Heather não é o que possui maior grau. Mas ele serve de ponte para os *nós* Ike e Jane, que estão alocados além dele, no restante da rede, o que faz com que assuma alto grau de importância. Por exemplo, se uma informação for passada de pessoa a pessoa, neste exemplo, Heather será de vital importância para garantir que Ike e Jane se comuniquem com o restante do grupo. Por outro lado, se ela for removida da rede, Ike e Jane não poderão se comunicar com o restante do grupo, o que faz com que Heather tenha alta centralidade *betweenness*.

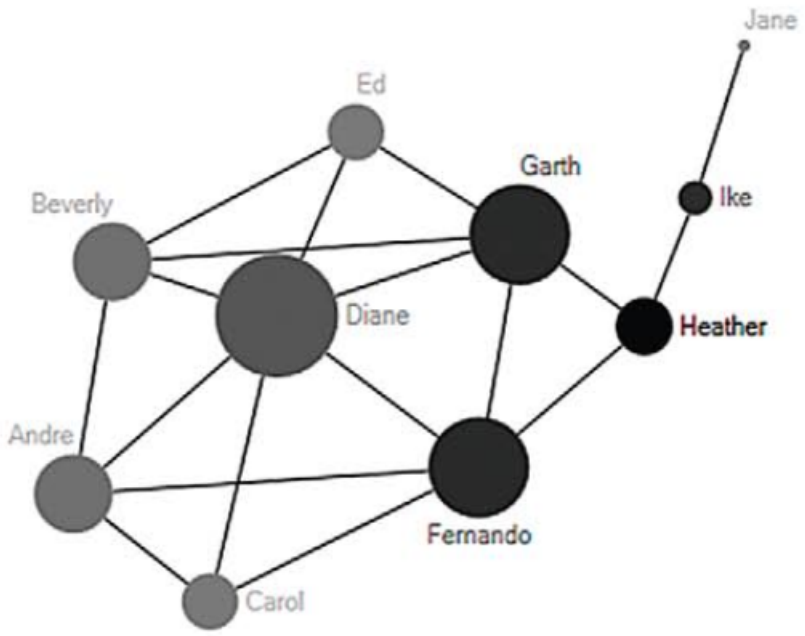


Figura 7: Exemplo de Centralidade por Intermediação.

Fonte: (HANSEN, 2010).

2.4.4 PageRank

Milhares de buscas são realizadas na Web diariamente por usuários procurando recuperar informações de seu interesse. Como retornar informações em ordem de relevância ou importância para o usuário? *PageRank* é uma medida utilizada pela *Google* como estratégia para classificar páginas existentes na *World Wide Web* de acordo com sua importância (PAGE, 1999). Newman (NEWMAN, 2018) alega que esta medida é parte central da tecnologia utilizada no motor de busca da empresa, classificando páginas de acordo com o real interesse de usuários (PAGE, 1999) (ZINOVIEV, 2018). Isto acontece a partir da estrutura de *links* da *Web*, como mostra a Figura 8 na qual páginas apontam e são apontadas por outras páginas. No trabalho de (ROGERS, 2002), o autor define *PageRank* como um 'voto' dado a uma página por outras páginas mostrando o quão importante ela é. Assim, além de contabilizar os 'votos' dados por outras páginas, também é contabilizado a importância de quem 'votou'. Esta relação é exibida na Figura 8, retirada de (345KAI, 2007), quando observado que C possui apenas o 'voto' de B, mas dada a importância de B, seu *PageRank* é mais alto que, por exemplo, E que possui mais 'votos', embora de páginas com baixa importância.

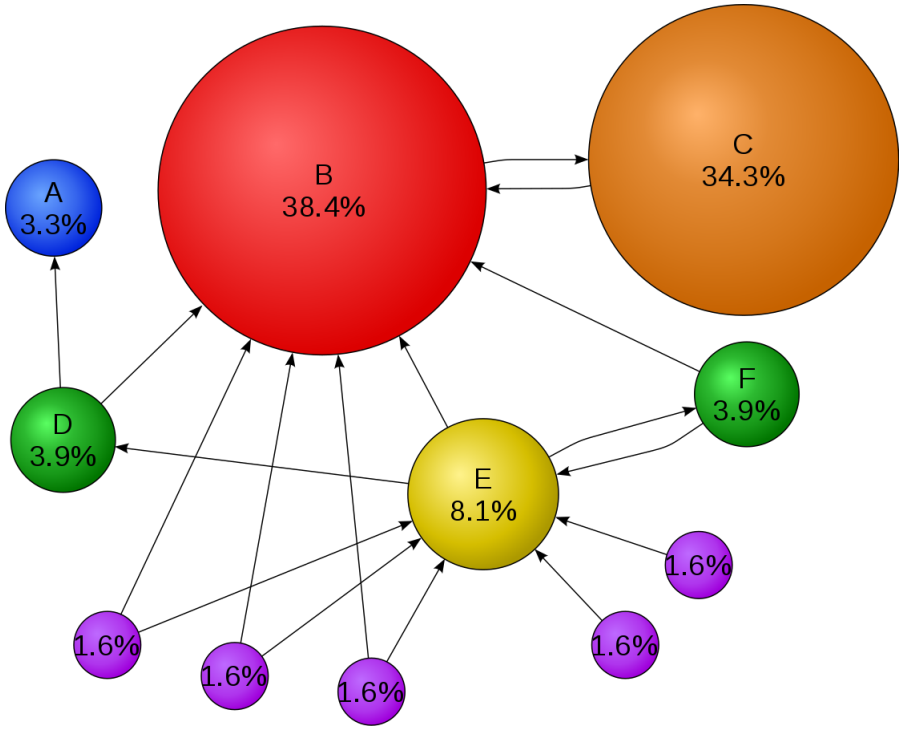


Figura 8: Exemplo de classificação de páginas *Web* de acordo com *PageRank*.

Fonte: (345KAI, 2007).

Por definição, o *PageRank* em Brin e Page (BRIN; PAGE, 1998) pode ser considerado como um modelo da interação do usuário na Web. De modo geral, um usuário utiliza os *links* disponibilizados na página atual, ou ainda, visita outra página via uma nova URL, entre outras possibilidades. A medida é apresentada em Page *et al* (PAGE, 1999), como uma versão simplificada, na equação:

$$R\left(u\right) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

Onde, $R_(u)$ é o *nó* para o qual será calculado o *PageRank*, B_u é o conjunto de arestas que apontam para u, R_u é o *PageRank* de v e N_v é o grau de v. Por fim, em Page *et al* é definido c como um fator de normalização que é a probabilidade de o usuário continuar seguindo *links* até a página desejada. O valor de c está definido em Brin e Page (BRIN; PAGE, 1998) como 0.85. Assim, a página com o *PageRank* mais alto é a mais atraente: não importa onde a pessoa comece, essa página é o destino final mais provável (ZINOVIEV, 2018).

Todas estas medidas auxiliam na compreensão da estrutura de uma rede. É intuitivo que quando, p. e., um problema é grande, a melhor forma de gerenciá-lo é dividindo em pequenas partes. No contexto de grafos, para (NEWMAN, 2010), seu particionamento possui como meta dividi-lo em pequenas partes, buscando a melhor divisão para que se possa gerenciá-las facilmente. Já a detecção de comunidade é utilizada como fer-

ramenta para esclarecer padrões de conexões, ou seja, o entendimento das estruturas de redes (NEWMAN, 2010), bem como extrair informações úteis sobre elas (SOBOLEVSKY, 2014). Redes de larga escala são um de um contexto onde essas metodologias podem ser melhor utilizadas, pois seus padrões de conexão podem não ser facilmente visíveis em sua topologia bruta (NEWMAN, 2010), diferente de partes dela que podem tornar sua exploração mais eficaz (GIRVAN; NEWMAN, 2002).

3 Detecção de Comunidades

De acordo com Lancichinetti e Fortunato, (LANCICHINETTI; FORTUNATO, 2009), sistemas complexos são categorizados por sua divisão em subsistemas de maneira hierárquica. Ou seja, sistemas complexos são compostos por partes interconectadas (ESTRADA, 2012), que exibem estrutura de comunidade, também considerados agrupamentos ou módulos (FORTUNATO; HRIC, 2016). A Figura 9, retirada de (GIRVAN; NEWMAN, 2002), apresenta uma rede de colaboração entre cientistas que trabalham no Instituto de Santa Fé (Santa Fé Institute) (NEWMAN, 2010) (FORTUNATO; HRIC, 2016). Neste caso, vértices/*nós* são cientistas e as arestas/*links* as coautorias. A rede possui agrupamentos bem definidos ou "estrutura de comunidade", presumivelmente refletindo divisões de interesses e grupos de pesquisa dentro do departamento. Observa-se ainda, *nós* que estão na fronteira de seu grupo. Estes podem agir como intermediários entre grupos, o que faz com que, representem um importante papel na dinâmica de processos de difusão através da rede (FORTUNATO; HRIC, 2016). (NEWMAN, 2010) (BARABASI, 2016) (JORA; CHIRA, 2016), (ZHANG, 2007) entre outros, compartilham desta visão de estrutura de comunidade.

Esta visão aparece em vários domínios da interdisciplinaridade da ciência de redes (LANCICHINETTI; FORTUNATO, 2009), tais como, estudo de interações entre diferentes grupos de pessoas ou animais, redes metabólicas, cujas comunidades são baseadas em grupos funcionais (JORA; CHIRA, 2016), interação de proteínas com função similar em redes proteicas, amigos em redes sociais, redes de citação (JORA; CHIRA, 2016) (FORTUNATO; HRIC, 2016) *websites* com tópicos similares na web (FORTUNATO; HRIC, 2016), Sociologia (NEWMAN, 2010) (MALLIAROS; VAZIRGIANNIS, 2013).

O problema de detecção de comunidades tem estado sob intensa investigação nos últimos anos (LANCICHINETTI; FORTUNATO, 2009) (ZHANG, 2007) (JORA; CHIRA, 2016). Assim, têm surgido métodos e ferramentas de diversas disciplinas (física, biologia, matemática aplicada, ciência da computação e ciências sociais) na busca pela solução, (LANCICHINETTI; FORTUNATO, 2009), bem com a utilização de diferentes abordagens (JORA; CHIRA, 2016). Nestes métodos e abordagens deve-se considerar que não existe restrições a tamanho de grupos como acontece em particionamento de grafos (BARABASI, 2016).

3.1 Agrupamento Hierárquico

Este método é considerado em (GIRVAN; NEWMAN, 2002) como um método tradicional para detectar estruturas de comunidade em redes. Tal como ilustra a Figura 10,

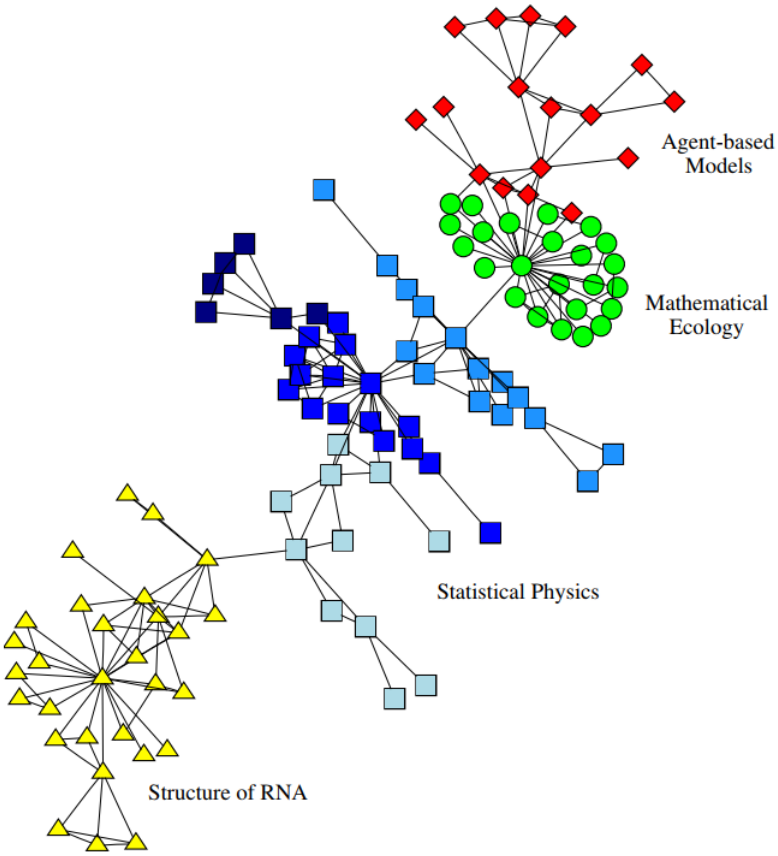


Figura 9: Rede de colaboração de cientistas atuando no Instituto de Santa Fé.

Fonte: (GIRVAN; NEWMAN, 2002).

resultado da execução do algoritmo dos autores, onde são encontradas três comunidades de nós densamente conectados e com baixa densidade de conexões entre elas.

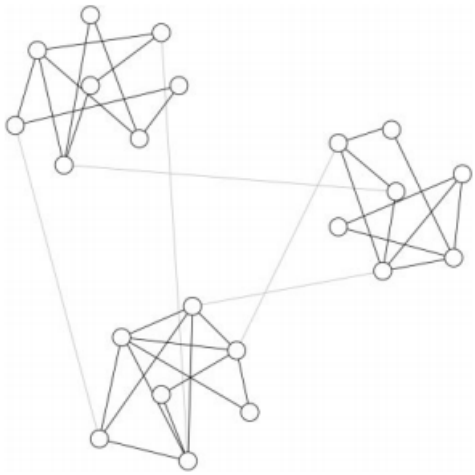


Figura 10: Esquema de representação de estrutura de comunidade baseado em agrupamento hierárquico gerado pelo método de Girvan e Newman.

Fonte: (GIRVAN; NEWMAN, 2002).

O método possui como ponto inicial, a matriz de similaridade , cujos elementos x_{ij} indicam a distância, do *nó* i para o *nó* j, neste caso a similaridade é extraída da posição relativa dos *nós* i e j dentro da rede (BARABASI, 2016). Ou seja, representa um certo sentido de quão estreitamente conectados os vértices estão (GIRVAN; NEWMAN, 2002). A partir de x_{ij}, o agrupamento hierárquico acontece com a identificação de grupos com alta similaridade. Há dois procedimentos para alcançar o agrupamento: algoritmos aglomerativos e divisivos. O primeiro incorpora *nós* com alta similaridade dentro da mesma comunidade, o segundo isola comunidades removendo arestas com baixa similaridade, que tendem a conectá-las (BARABASI, 2016). Ambos os métodos geram uma árvore hierárquica (BARABASI, 2016) (GIRVAN; NEWMAN, 2002) (ZHANG, 2007) também chamado de dendrograma[1], como mostra a Figura 11, que ilustra possíveis partições de comunidade. O nível mais baixo no qual dois nós estão conectados representa o força da aresta x_{ij} e a ordem na qual os pares se juntam para formar uma comunidade (GIRVAN; NEWMAN, 2002).

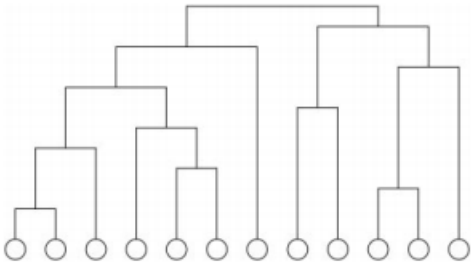


Figura 11: Exemplo de um pequeno dendrograma de agrupamento hierárquico.

Fonte: (GIRVAN; NEWMAN, 2002).

O funcionamento dos métodos aglomerativos e divisivos podem ser verificados em algoritmos propostos por Ravasz (RAVASZ, 2002) e Girvan e Newman (NEWMAN; GIRVAN, 2004), respectivamente.

Algoritmos para detectar comunidades sempre produzem alguma divisão, mesmo para uma rede aleatória que não possui estrutura de comunidade significativa (NEWMAN; GIRVAN, 2004). A questão aqui é: como medir a qualidade de uma partição?

3.1.1 Medida de Qualidade de uma Partição

A questão está em avaliar, de possíveis partições obtidas de uma rede, aquela que é a melhor. Assim, é definida uma medida de qualidade, Q modularidade (NEWMAN, 2006). A ideia principal está baseada na definição de um subconjunto de *nós* como comunidades se as conexões internas entre seus *nós* é em número maior que o esperado

[1] dendrograma - um diagrama de ramificação que representa uma hierarquia de categorias com base no grau de similaridade ou número de características compartilhadas.

se este subconjunto tivesse uma estrutura formada aleatoriamente (XAVIER, 2015). Em Newman (NEWMAN, 2010), quanto maior o valor desta medida melhor é a divisão da rede, desconsiderando em quantos grupos ela se divide. A ideia é que uma estrutura de comunidade verdadeira em uma rede está relacionada ao arranjo estatístico das arestas e pode ser quantificado com esta medida (NEWMAN, 2006). O conceito de modularidade em (RAVASZ, 2002), assume que funcionalidade celular pode ser corretamente particionada dentro de uma coleção de módulos. Cada módulo é uma entidade discreta de vários componentes e executa uma atividade identificável, separada das funções de outros módulos. Como em particionamento de grafos, maximizar a modularidade é um problema difícil (NEWMAN, 2010). Neste caso, podem ser utilizados algoritmos heurísticos que buscam maximizar esta medida de forma inteligente provendo, na maior parte do tempo, bons resultados (NEWMAN, 2010), visto que algoritmos não heurísticos levam tempo exponencial para executar a divisão com máxima modularidade (NEWMAN; GIRVAN, 2004) (NEWMAN, 2010) (BARABASI, 2016). Valores de modularidade podem ser positivos e negativos. Quando positivos, indicam que possivelmente é encontrada estrutura de comunidade (NEWMAN, 2006) (BARABÁSI, 2011). Valores entre 0.3 e 0.7 são mais comuns de serem encontrados, por outro lado, valores acima deste intervalo são raros (NEWMAN; GIRVAN, 2004). Assim, para Newman e Girvan (NEWMAN; GIRVAN, 2004) a modularidade Q avalia a qualidade de uma partição considerando a fração de arestas dentro da mesma comunidade com o número esperado de uma formação aleatória. A modularidade pode ser descrita como:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta c_i c_j,$$

onde A_{ij} são os elementos da matriz de adjacência de $G(V, E)$; k_i é o grau de saída do *nó* i; $m = |E|$ e $\delta c_i c_j$, delta de Kronecker, é 1 se i e j pertencem a mesma comunidade e 0, caso contrário. Neste contexto, não é considerado a direção de arestas. Todavia, em (LEICHT; NEWMAN, 2008), os autores, descrevem um algoritmo que aborda qualidade de partições em redes direcionadas.

Um exemplo retirado de (BARABASI, 2016), diz que a modularidade se dá da seguinte forma: considere uma rede com N *nós* e L *links* e uma partição que contém n_c comunidades. Neste caso, cada comunidade possui N_c nós conectados por L_c arestas. Para que um subgrafo possa ser considerado uma verdadeira comunidade, o número de L_c deve ser mais alto internamente do que entre grupos N_c, obedecendo a sequência de graus da rede. De maneira que um subgrafo C_C poderia fazer parte de uma comunidade real, como exibido na Figura 12, retirada do trabalho do citado autor.

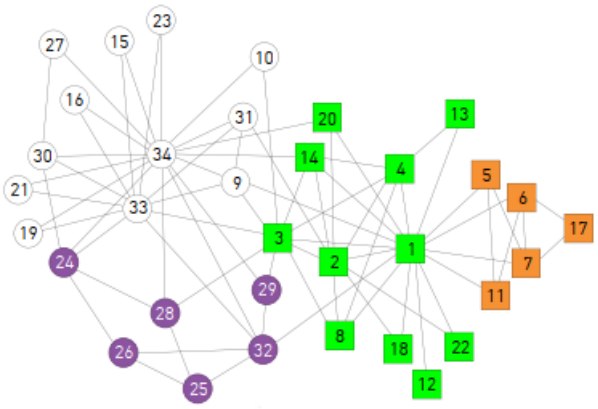


Figura 12: Particionamento realizado pelo algoritmo guloso de Newman, exibindo a estrutura de comunidade da rede Karate Clube.

Fonte: (BARABASI, 2016).

Ao longo do tempo, conjuntos de redes têm se expandido e assim, a importância da velocidade dos algoritmos de detecção de comunidades se torna maior (GREGORY, 2010). O algoritmo de Propagação de Rótulos (Label Propagation Algorithm) é considerado pelos autores (RAGHAVAN, 2007), bem como por (LIU; MURATA, 2010) e (GREGORY, 2010), um dos algoritmos mais rápidos para este processo. Neste trabalho, os autores (RAGHAVAN, 2007) propõem apenas a utilização da estrutura da rede, não direcionada, para orientar o processo de detecção de comunidades sem especificar o número e o tamanho das comunidades, que são definidas ao final da execução. O algoritmo inicializa atribuindo a cada *nó* um rótulo único e a cada iteração do algoritmo os *nós* atualizam seu rótulo de acordo com o rótulo da maioria de seus vizinhos, quebrando laços de maneira uniforme e aleatória. Com isso, diante do consenso na escolha dos rótulos, formam grupos de *nós* densamente conectados. No final do algoritmo, os *nós* com os mesmos rótulos são agrupados em comunidades. O autor considera como vantagens em relação a outros métodos, a simplicidade e eficiência de tempo. O algoritmo é validado com sua utilização nas redes de amizade 'Karate Club' e 'U.S. college football', visto que o resultado está de acordo, em ambas as redes, com a estrutura de comunidade presente. A relevância do algoritmo é encontrada na literatura e é garantida por algoritmos que dele se utilizam como base para criação de novos algoritmos de detecção de comunidade. Tais como; (GREGORY, 2010) com o algoritmo COPRA, (COSCIA, 2012) com DEMON, (XIE, 2011) com SLPA, (LIU; MURATA, 2010) com o LPAm, LPAm+.

Todavia, observando a Figura 12, a exibição de comunidades apresenta conjuntos disjuntos. Embora, redes do mundo real apresentem elementos que participam em mais de uma comunidade simultaneamente.

3.2 Detecção de Comunidades com Sobreposição

Sistemas complexos da natureza e sociedade, de acordo com Palla *et al* (PALLA, 2005), podem ser descritos em termos de redes que capturam a complexidade de suas conexões entre as unidades que os constituem. Neste trabalho, os autores, levantam a questão de como interpretar a organização global de tais redes em comunidades de maneira que se possa entender as propriedades estruturais e funcionais da rede. Entretanto, métodos determinísticos usados para encontrar comunidades disjuntas em grandes redes desconsidera que elas são arranjos de grupos *nós* coesos altamente sobrepostos. P. e., em (BARABÁSI, 2011) um mesmo gene que é suspeito estar em múltiplas doenças é uma indicação de que ele está na sobreposição de uma "rede de doenças" corroborando o estudo de detecção de comunidades sobrepostas em distintos domínios de pesquisa (LANCICHINETTI, 2009) (NEWMAN, 2010) (GREGORY, 2010) (NICOSIA, 2009). A Figura 13, retirada de (FORTUNATO; HRIC, 2016), exibe a visão clássica da estrutura de comunidades, onde subgrafos são densamente conectados. Por outro lado (PALLA, 2005), exemplifica o contexto de comunidades sobrepostas da seguinte forma: um cientista, que além da comunidade cientifica com a qual compartilha seus interesses profissionais, também pertence a uma comunidade que suporta membros da família e parentes, bem como, talvez, outra comunidade de indivíduos que compartilham algum *hobby*. A Figura 14, ilustra bem este conceito de comunidade sobreposta. Na Figura 14a um indivíduo da rede, o ponto preto central, configura um indivíduo que pertence a várias comunidades. A imagem detalha a estrutura aninhada de comunidades sobrepostas, p. e., na comunidade científica, ainda estão definidas a hierarquia de comunidades definidas pelos subconjuntos (cientistas da Física, da Matemática, da Biologia e do Departamento de Física Biológica). Modela, também, de forma abstrata, a comunidade da família, retratando a cascata de comunidades iniciando deste indivíduo central e exemplificando a estrutura entrelaçada da rede de comunidades. Na Figura 14b os métodos aglomerativos e divisivos falham em identificar comunidades sobrepostas significantes. E, por fim, a Figura 14c a sobreposição de quatro de comunidades, nesse exemplo, a comunidade amarela se sobrepõe a azul em apenas um nó, por sua vez, ela compartilha coma a comunidade verde dois *nós* e um *link*. Ou seja, a maioria das redes atuais é composta por grupos coesos e de nós altamente sobrepostos.

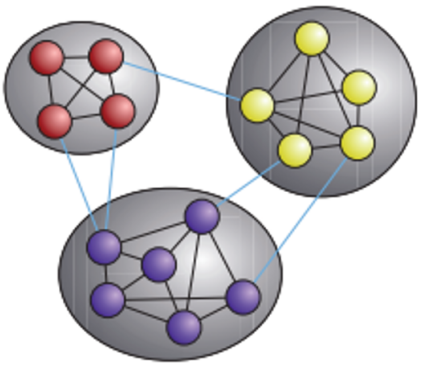


Figura 13: Visão clássica da estrutura de comunidade.

Fonte: (FORTUNATO; HRIC, 2016).

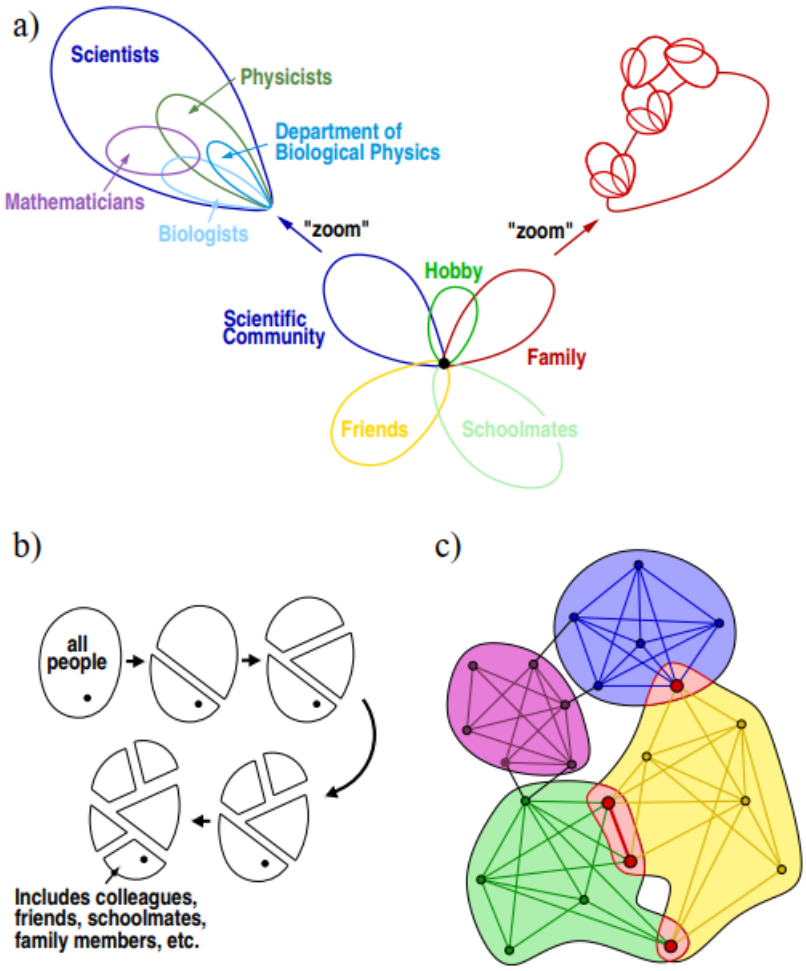


Figura 14: Ilustra o conceito de comunidades sobrepostas.

Fonte: (PALLA, 2005).

Redes sociais são um exemplo claro de comunidades sobrepostas, pois é comum indivíduos estarem em diferentes círculos ao mesmo tempo, tais como: amigos, família, trabalho, entre outros interesses (NEWMAN, 2018). Encontrar comunidades em grandes

redes supõe a existência de métodos que sejam úteis se a estrutura de comunidade pode ser interpretada em termos de conjuntos separados, como ilustrado na Figura 14(b) (PALLA, 2005).

Palla et al (PALLA, 2005) foram os primeiros a trazer esta questão para a comunidade da ciência de redes. Neste caso, os autores propõem uma abordagem para analisar os principais recursos estatísticos dos conjuntos sobrepostos de comunidades. Neste sentido, comunidade para eles consiste de vários subgrafos completamente conectados, que possuem a tendência a compartilhar alguns *nós*. Definição que busca representar uma característica essencial de uma comunidade, o fato de que seus membros possam ser alcançados por meio de subconjuntos de nós bem conectados.

Posteriormente, no trabalho de (ADAMCSEK, 2006), é introduzido um método para localizar e visualizar comunidades sobrepostas, CFinder. O algoritmo de busca utiliza o CPM (Clique Percolation Method) que é descrito em (DERÉNYI, 2005). Ele inicia com a identificação de todos os cliques de tamanho k na rede, a seguir, uma novo grafo é construído; de maneira que cada *nó* representa um *k-clique*. Assim, dois *nós* estão conectados se o *k-clique* que o representa compartilha $k - 1$ membros.

3.2.1 Algoritmos de Detecção de Comunidades baseados no LPA

Os autores de (COSCIA, 2012) consideram que é mais profícuo focar primeiramente na estrutura local de comunidades, para só assim encontrar a estrutura a nível global. Dessa forma, o algoritmo DEMON (*Democratic Estimate of the Modular Organization of a Network*) propõe uma abordagem simples local-first (local primeiro) para descoberta em redes complexas que seja capaz de desvendar a estrutura modular que emerge de padrões locais (COSCIA, 2012). O intuito é alcançado quando se permite a um *nó* (considerado o nó ego da rede), democraticamente, votar nas comunidade; dada sua visão limitada do seu entorno. Um algoritmo de propagação de rótulos (LPA) é utilizado para este processo e, assim, comunidades locais são mescladas em um conjunto global.

O algoritmo SLPA (*Spekers-Lisener Label Propagration*), diferente do LPA, para o qual cada *nó* tem atribuído um rótulo único e, assim, é iterativamente atualizado pela adoção do rótulo da maioria de seus vizinhos. Como já visto, este processo encontra comunidades disjuntas quando o algoritmo converge. O algoritmo proposto pelos autores (XIE, 2011), segue a ideia de (GREGORY, 2010) de *nós* possuindo vários rótulos de forma a permitir a sobreposição. A dinâmica do processo do algoritmo SLPA precisa determinar: (i) como a informação de um *nó* espalha para outros; (ii) como processar informações recebidas de outros. Ambas as questões dizem respeito em como prover a manutenção da informação, de maneira que, a solução encontrada está a imitar o comportamento de comunicação humano, propagação da informação pelo processo de (*speakers-listener*). No processo, cada nó pode ser ouvinte ou falante, dependendo do papel do nó no momento,

fornecer ou consumir informação. Em geral, um *nó*, durante o processo, pode armazenar vários rótulos. Todavia, isso depende de como os processos estocásticos aconteceram, conduzidos pela estrutura de rede subjacente. Em vez de apagar o conhecimento de todos os rótulos observados, um *nó* os acumula repetidamente, apagando apenas um deles. Ademais, "quanto mais um nó observar um rótulo, maior a probabilidade de espalhar esse rótulo para outros *nós* (imitando a preferência das pessoas de espalhar opiniões com mais frequência)".

3.2.2 Avaliando Partições de Comunidades com Sobreposição

Com a disponibilidade crescente de algoritmos que buscam encontrar a estrutura de comunidades em redes do mundo real, e assim, estruturas de comunidades sobrepostas, também se torna essencial que estas partições sejam avaliadas. De maneira que, cresce a demanda em definir e medir de alguma forma as diferentes partições fornecidas por vários métodos (LÁZÁR, 2010). O trabalho de quantificar a qualidade de uma dada partição de rede com comunidades sobrepostas é mais difícil do que para comunidades que não sobrepõem, dado que um *nó* pode pertencer a várias comunidades ao mesmo tempo, porém sua força de conexão para cada uma delas é diferente (NICOSIA, 2009).

No trabalho de (NICOSIA, 2009), os autores, estendem o conceito de modularidade de (NEWMAN; GIRVAN, 2004) para redes sobrepostas, considerando o nível de 'força' com a qual um dado *nó* i pertence *k-ésima* comunidades. Neste caso, um conjunto de comunidades sobrepostas definidas a partir de um grafo direcionado $G(E, V)$ e uma matriz de "fatores de pertencimento" $\left[\alpha_{i,1}, \alpha_{i,2}, ..., \alpha_{i,|k|}\right]$ poderá ser atribuída a cada *nó* i, de maneira que cada coeficiente $\alpha_{i,k}$ define a força com a qual este *nó* pertence a comunidade k. Os autores ainda afirmam que podem, sem perda de generalidade exigir que:

$$0 \leq \alpha_{i,k} \leq 1 \qquad \forall i \in V, \qquad \forall k \in K$$

e que,

$$\sum_{k=1}^{|K|} \alpha_{i,k} = 1.$$

Dessa forma, para cada aresta l_{il}, conectando o *nó* i e j na comunidade k é atribuído um coeficiente de pertencimento, chamado de β_{lk}. Definido como:

$$\beta_{l,k} = F\left(\alpha_{i,k}, \alpha_{j,k}\right),$$

de maneira que os autores definem modularidade, para comunidades com sobreposição como:

$$Q_{ov} = \frac{1}{m} \sum_{k \in K} \sum_{i,j \in V} \left[\beta_{l(i,j)k} A_{ij} - \frac{\beta^{out}_{l(i,j),k} K^{out}_i \beta^{in}_{l(i,j),k} K^{in}_j}{m} \right].$$

Onde:

- $\beta_{l(i,j)k} A_{ij}$ é o coeficiente de pertencimento de uma aresta i e j para uma comunidade k;
- $\beta^{out}_{l(i,j),k} K^{out}_i \beta^{in}_{l(i,j),k} K^{in}_j$ é a média de todos os coeficientes de pertencimento de uma aresta l para uma comunidade k;
- m é o número de arestas da comunidade.

Todo levantamento teórico, visto até agora, dimensiona conectividade em uma rede a nível estrutural. Quem está conectado a quem, bem como, sua importância do ponto de vista de algumas métricas. O Capítulo 4 levantará questões relacionadas a conectividade a nível de comportamento.

4 Difusão em Redes

Este Capítulo busca definir alguns conceitos sobre difusão e dinâmicas do processo, visto que a proposta deste trabalho é avaliar a difusão em um contexto social.

Empiricamente, pode-se observar indivíduos na sociedade e como, em alguns casos, o comportamento de alguns dita a conduta de outros. Esse 'movimento', denominado, comportamento coletivo em (GRANOVETTER, 1978), está baseado em situações onde um indivíduo possui duas opções de escolha, ou seja, qual o custo ou benefício de uma decisão dependendo do comportamento de outros. O autor afirma que o conceito chave "*threshold* (limite) diz respeito ao número ou a proporção de outros que devem tomar uma decisão antes que uma determinada pessoa o faça; este é o ponto em que os benefícios líquidos começam a exceder os custos líquidos para essa pessoa em particular".

Deste contexto, observa-se outros dois conceitos; influência social e difusão da inovação. O primeiro, segundo Rashotte(RASHOTTE, 2007), diz que "influência social é definida como mudanças em um indivíduo, de pensamento, sentimento e, atitudes que resultam da interação com outros indivíduos ou grupos" . O segundo, difusão de inovação, de acordo com Rogers (ROGERS, 2010) é um processo social no qual o comportamento de um indivíduo esta intimamente relacionado ao comportamento de seus contatos.

Para corroborar com os conceitos acima, no trabalho de (EASLEY, 2010), difusão de inovação é utilizado na sociologia na construção de processos de difusão onde são considerados especialmente, como novos comportamentos, práticas, ideias, opiniões, convenções e tecnologias fluem de pessoa para pessoa em uma rede social.

O autor apresenta um modelo bem básico da difusão de um novo comportamento considerando o conceito de jogo de coordenação. O exemplo é colocado de forma que em uma rede social, cada *nó*/jogador tem a possibilidade de escolher entre dois comportamentos, A e B, se o *nó* u e v estão ligados, existe a possibilidade de compartilharem o mesmo comportamento. O jogo é iniciado com u e v, as estratégias A e B e a recompensa é definida da seguinte forma:

- Se a estratégia A for adotada por ambos os jogadores, ambos possuem recompensa $a > 0$;
- Se ambos adotam B, cada um possui recompensa $b > 0$;
- Se ambos divergem na escolha da estratégia, ambos tem recompensa de 0.

A Figura 15 demonstra este conceito, retirado de (EASLEY, 2010), como uma matriz de recompensa.

		w	
		A	B
v	A	a, a	$0, 0$
	B	$0, 0$	b, b

Figura 15: A-B Jogo de Coordenação.

Fonte: (EASLEY, 2010).

A colocação acima avalia apenas uma tomada de decisão individual, ou seja, a existência de uma única aresta. Em uma rede, como um todo, esse exemplo explora o comportamento em cascata (EASLEY, 2010). Este fenômeno ocorre quando um indivíduo toma uma decisão que se mostra ótima observando as ações daqueles que o rodeiam (BIKHCHANDANI, 1992). Supondo uma rede onde todos estão usando B por padrão, e apenas alguns poucos escolham utilizar A, por acreditarem obter uma recompensa maior. De maneira que alguns de seus vizinhos decidam segui-los na escolha, e os vizinhos deles, formando uma potencial cascata (EASLEY, 2010). Neste ponto os autores levantam duas questões: todos adotam A?; senão, quando a difusão de A para?

Para (EASLEY, 2010), a estrutura da rede responde a essa questão, bem como a escolha dos adotadores iniciais e ainda, o valor de *threshold* (limite) que cada *nó* utiliza para decidir se escolhe a estratégia A. (LÓPEZ-PINTADO, 2008) mais especificamente coloca a importância da distribuição de conectividade na rede. De maneira que, a partir da estrutura da rede, observa-se como indivíduos são influenciados por seus vizinhos específicos. Esta dinâmica é corroborada por Gullati (GULATI, 1999) e Tidd (TIDD; BESSANT, 2018) quando afirmam que uma rede pode influenciar ações de indivíduos de duas formas. A primeira, por meio do fluxo de informação na rede e, segundo, por meio da localização dos atores na mesma.

Outra consideração de (EASLEY, 2010) é a ocorrência da quebra de um processo de difusão de um novo comportamento devido a tentativa de penetrar em uma comunidade densamente conectada dentro da rede. Neste caso, os autores, consideram que a homofilia pode servir como uma barreira para a difusão.

Considere um grupo de amigos enclausurados dentro de um grupo com interesses comuns. Alguns resolvem adotar uma nova tecnologia, pressupondo que seu uso seja vantajoso para o grupo. Possivelmente outros amigos irão utilizar. Todavia, outros grupos podem oferecer resistência a esta adoção proposta.

4.1 Maximização de Influência

Uma das questões fundamentais encontrada no contexto de difusão é o problema de maximização de influência, ou seja, deseja-se maximizar o número de pessoas adotando

um determinado produto, crença, ideia, serviço, tecnologia, etc., onde o limite de recursos é um fator a ser considerado. Um exemplo simples, fornecido por (KEMPE, 2003), seria o de ofertar a adotantes iniciais um produto para que possam utilizar e indicar aos amigos; esperando que ao final de um determinado intervalo de tempo, seja obtido um maior número de usuários do produto (ZHANG, 2014).

Em 2001, Domingos e Richardson (DOMINGOS; RICHARDSON, 2001), modelaram o potencial valor de um consumidor não apenas em seu poder de compra de um determinado produto, mas em seu poder de influenciar outros na compra deste produto. Os autores definem o conceito de marketing direto como a tentativa de selecionar clientes para que utilizem um produto e possuam maior probabilidade de influenciarem outros a utilizarem. Quanto maior o alcance da influência, maior o valor destes clientes, tornando-os mais lucrativos. Assim, é avaliada uma rede de clientes, onde observa-se destes quais são mais profícuos em influenciar outros, e estes outros mais outros, sucessivamente.

Em (KEMPE, 2003), os autores, motivados pelo trabalho de (DOMINGOS; RICHARDSON, 2001), propuseram um problema algoritmo para influência em redes sociais. Tal problema possui o seguinte questionamento: dado um subconjunto de indivíduos, como convencê-los a adotar um novo produto ou inovação de maneira a que se alcance uma grande cascata de adoção, qual conjunto de indivíduos escolher? O problema é formalizado como um problema de otimização discreta, cuja complexidade se mostra *NP-hard* para muitos dos modelos presentes na literatura, incluindo *Threshold Model* e *Independent Cascade Model*. Os autores expandem o tema em mais dois artigos (KEMPE, 2005) e (KEMPE, 2015).

Formalmente, o modelo é definido em (CHEN, 2010) como: dada uma rede social modelada como um grafo não direcionado $G = (V, E)$, onde os indivíduos são modelados a partir de V e o relacionamento entre eles, a partir de E. Neste caso, espera-se que a partir de um pequeno conjunto de sementes, indivíduos considerados hábeis em promover a propagação de um processo de difusão, fazendo com que o comportamento dos mesmos se transforme em uma cascata na rede (KEMPE, 2003). Para Kempe *et al* (KEMPE, 2003), a influência esperada está relacionada a seu processo de propagação que é capturado pelos modelos de difusão, os quais estarão presentes na Seção 4.2.

4.1.1 Trabalhos Relacionados

No trabalho de (KEMPE, 2003), o problema de maximização de influência pode ser aproximado utilizando como estratégia um algoritmo guloso simples. Este tipo de algoritmo obtém um conjunto de sementes selecionando repetidamente *nós* que fornecem um aumento marginal de influência a cada passo de tempo. Neste caso, a garantia da execução está na propriedade submodular da função da influência arbitrária $f(.)$, a qual é definida se o ganho em adicionar um elemento a um conjunto S é menor do que o ganho

em adicionar o mesmo elemento a um conjunto T que contenha S. A função submodular satisfaz

$$f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T),$$

para todos os elementos v e todos os pares de $S \subseteq T$. Os autores mostram que a estratégia gulosa obtém aqui desempenho melhor, superior a 63%, do que a estratégia de (DOMINGOS; RICHARDSON, 2001). A abordagem proposta por (KEMPE, 2003) fornece resultados que estão dentro do ótimo para várias classes de modelos de difusão.

De (LESKOVEC, 2007), o algoritmo CELF *(Cost-Effective Lazy Forward)* generaliza o trabalho de (KEMPE, 2003) e explora a propriedade de submodularidade para obter uma solução eficiente em selecionar *nós* influentes por meio de um algoritmo guloso. Sua proposta considera que *nós* possuem custos ligados a eles que não são constantes, ou seja, cada qual possui um custo e, neste caso é necessário obter na escolha aqueles, cujo custo fornecem maior benefício. A função utilizada para a escolha de *nós* mais influentes é chamada *Lazy-Forward*. A ideia é que para uma função conjunto submodular, utilize-se a conhecida técnica *lazy evaluations*, cujo objetivo é diminuir significativamente o número de avaliações sem, contudo, modificar o resultado do algoritmo guloso (LESKOVEC, 2007) (CHEN, 2013). O processo de execução de *lazy evaluations* é manter uma fila com o ganho marginal de cada *nó* em ordem decrescente. Após a adição de um determinado *nó* no conjunto semente, todos os outros na fila de prioridades são marcados como nulos. Para alocar um outro elemento no conjunto de sementes, recalcula-se o benefício do *nó* do início da fila e o insere na posição adequada. Em muitos casos, o valor recalculado é pouco menor que o valor anterior e assim este *nó* é adicionado ao conjunto de sementes. O algoritmo dos autores, se adapta a grandes problemas e alcança uma melhor solução por meio da estratégia gulosa do que o algoritmo de (DOMINGOS; RICHARDSON, 2001). O mesmo provou alcançar resultados próximos do ótimo e, até 700 vezes mais rápido que um algoritmo guloso simples.

Em (CHEN, 2009) é proposto também uma melhoria para a execução de um algoritmo guloso. Os autores a partir de um conjunto S selecionado para iniciar a propagação de influência, aplicam a este conjunto um processo aleatório, *RanCas(S)*, o qual retornará um conjunto aleatório de vértices influenciados por S. O algoritmo, no trabalho, considera uma rede G e um número K como entrada a fim de gerar o conjunto inicial de sementes com número de elementos K. Neste caso, espera-se que o resultado da cascata de influência seja o maior possível. Para melhor compreensão do processo, veja o trabalho dos autores. Além da melhoria proposta para execução do algoritmo guloso, os autores também propõem uma heurística, *degree discount*, que se baseia no grau de um *nó*, a qual é focada para o modelo *Independent Cascade*. A utilização está fundamentada na literatura da sociologia, onde heurísticas de grau, bem como outras baseadas em centralidade são utilizadas para estimar a influência de *nós* em redes sociais, como alega (WASSER-

MAN, 1994). A ideia da utilização da heurística, *degree discount* é que quando um *nó* u é selecionado para o conjunto de sementes, considerando seu valor de grau; se o próximo *nó* v a ser selecionado é vizinho de u, existe uma aresta entre (u, v), é realizado um desconto de um no grau de v. De maneira que, para o modelo *Independent Cascade* com pequena probabilidade de propagação, se o *nó* u já está no conjunto de semente, e provavelmente ele influenciará v, não há necessidade de selecionar v para conjunto de sementes. Os autores alegam que o algoritmo também supera heurísticas de seleção de *nós* que se baseiam em centralidade de grau e distância em redes sociais.

Em (GOYAL, 2011), os autores propõem uma melhoria para o algoritmo CELF de (LESKOVEC, 2007), CELF++, para o qual os autores empiricamente mostram que este alcança maior velocidade que seu percursor, na faixa de 35-55%. Os autores designam $\sigma(S)$ para denotar o conjunto de sementes para a difusão. Neste caso, atributos dos *nós* são armazenados em uma lista na forma de; $\langle u.mg1, u.prev_best, u.mg2, u.flag\rangle$. Assim, $u.mg1 = \Delta_u(S)$, define o ganho marginal de u para o atual conjunto de sementes; $u.prev_best$ é o *nó* que apresenta o maior ganho marginal entre todos examinados na atual iteração, antes do *nó* u; $u.mg2 = \Delta_u\left(S \cup \{prev_best\}\right)$ e, $u.flag$ define a iteração na qual $u.mg1$ foi atualizado. Isto suprime a necessidade de recalcular o ganho de $\Delta_u\left(S \cup \{prev_best\}\right)$ na próxima iteração se $u.best_prev$ foi adicionado ao conjunto de sementes na iteração atual. Adicionalmente, a avaliação de $\Delta_u(S)$ e $\Delta_u\left(S \cup \{prev_best\}\right)$ podem ser realizadas simultaneamente.

Do trabalho de Kempe *et al* (KEMPE, 2003) até os dias atuais, muito se tem feito na busca por melhorias significativas para o problema de maximização de influência, bem como trabalhos que buscam propiciar melhor compreensão sobre o tema.

Essa discussão leva ao estudo de modelos de difusão em redes sociais.

4.2 Modelos de Difusão

Modelos de difusão têm sido utilizados para identificar a taxa de influência, a proporção final de uma população influenciada e a direção de uma adoção de inovações (TIDD; BESSANT, 2018). Neste caso, a taxa de influência/adoção é a velocidade relativa com a qual uma inovação é adotada por seus participantes em um sistema social (ROGERS, 2010).

A difusão de uma inovação é para (ROGERS, 2010) (TIDD; BESSANT, 2018) descrita, geralmente, por uma curva na forma de S, como mostra a Figura 16 presenta no trabalho de (ROGERS, 2010); trata do estudo sobre adoção de milho hibrido em uma comunidades do estado de Iowa nos Estados Unidos. No início, a taxa de adoção é baixa devido ao fato do processo de adoção estar limitado aos inovadores. Aumentando, lentamente, com poucos adotantes ao longo do tempo e, acelerando quando metade dos

indivíduos do sistema adotam. A curva desacelera a partir do momento que restam pouco inovadores que adotem a inovação (ROGERS, 2010).

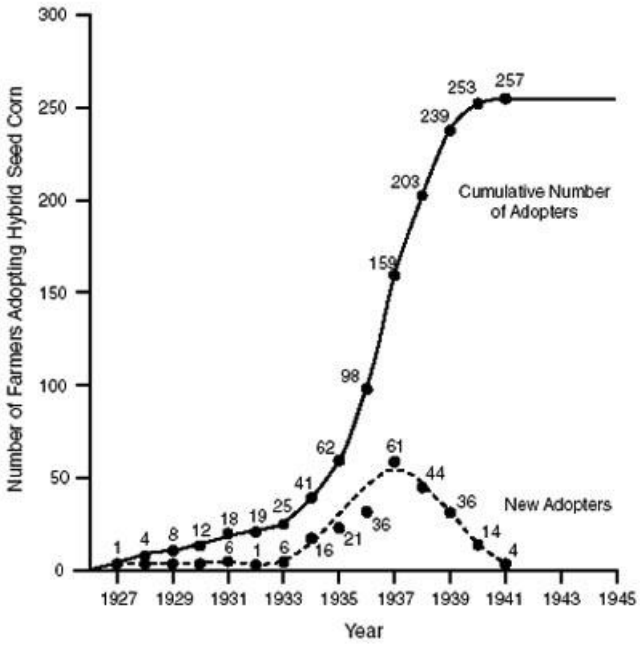


Figura 16: Processo de Difusão em forma de S.

Fonte: (ROGERS, 2010).

Para Zhang *et al* (ZHANG, 2014), os modelos de difusão podem ser categorizados dentro de três classes: *threshold models*, *cascading models* e *epidemic models* ; cada uma delas se desdobrando em múltiplas variações.

Além de (ZHANG, 2014), vários autores identificam o *Linear Threshold* e *Independent Cascading* como os mais populares dos modelos. De maneira que nas próximas Seções 4.2.1 e 4.2.2 são descritos os processo de difusão destes modelos, bem como condições de ativação.

4.2.1 Modelos *Threshold*

(ZHANG, 2014) evidencia como *Theshold Models* caracterizam comportamentos coletivos. Na modelagem matemática ou estatística, um modelo de *threshold* é qualquer modelo em que um valor de limite, ou conjunto de valores de limite, é usado para distinguir intervalos de valores em que o comportamento previsto pelo modelo varia de alguma maneira importante.

Segundo (GRANOVETTER, 1978), o modelo trata de decisões binárias, onde atores (*nós* da rede) possuem dois distintos e mutuamente exclusivos comportamentos alternativos, ou adota um novo comportamento ou não. Um exemplo desta situação é categorizada no trabalho de (EASLEY, 2010) com a seguinte questão: qual a probabilidade

de um indivíduo decidir se toma parte em um protesto contra o governo ou não? O autor coloca a situação da seguinte forma: organizar um protesto em uma sociedade sob regime repressivo. Imagine um indivíduo vivendo em tal sociedade e esteja ciente de uma manifestação pública contra o governo que está planejada para um determinado dia. Caso um número expressivo de pessoas aparecer, então o governo ficará seriamente enfraquecido e todos na sociedade, incluindo os manifestantes, irão se beneficiar. Mas se apenas algumas centenas aparecerem, então os manifestantes simplesmente serão presos (ou pior). De maneira que, teria sido melhor se todos tivessem ficado em casa. Nestas circunstâncias, qual o melhor comportamento o indivíduo deve adotar? Nesta situação, (GRANOVETTER, 1978), mostra que um outro requisito deste tipo de decisão é que o custo e benefício para que o ator tome uma decisão depende muito de quantos outros o fizeram. Dessa forma, no exemplo dado por (EASLEY, 2010), qual o custo e o benefício de se participar de um protesto contra o governo?

Os indivíduos nesse modelo (GRANOVETTER, 1978), são considerados racionais, eles agem de maneira que percebendo as circunstâncias, seus objetivos e preferências sejam maximizadas. O modelo foca na diferença entre os indivíduos, pois indivíduos diferentes exigem diferentes níveis de segurança para adotar um novo comportamento. O autor cita que esta variação entre os indivíduos é o limite, isto é, o limite para uma pessoa adotar um novo comportamento, produto ou ideia é a proporção dos que já adotaram um novo comportamento, antes que ele o faça.

O modelo do *Theshold model* pode ser aplicado, segundo Granovetter (GRANOVETTER, 1978) em tais situações: difusão de inovação, rumores e doenças, greve, eleições, promoção educacional, se ausentar de ocasiões sociais e psicologia social experimental. Para o autor, o objetivo o mesmo para todos esses cenários, predizer, a partir da distribuição inicial dos limites, o número máximo ou a proporção entre as duas decisões possíveis.

O modelo *Linear Threshold*, ou Limite Linear, trata as questões acima e é, das variações do modelo *threshold* considerado por (ZHANG, 2014) a mais amplamente estudada. Para o modelo em (SHAKARIAN, 2015) e (ZHANG, 2014), cada aresta direcionada $(u, v) \in E$ possui um peso não negativo $b(u, v)$ (função linear). Para qualquer *nó* $v \in V$, o total do *in-degree* soma menos que ou igual a um, ou seja, $\sum_{u \in \eta^{in}(v)} b(u, v) \leq 1$ (SHAKARIAN, 2015).

Dessa forma, o modelo é definido em (SHAKARIAN, 2015) e (ZHANG, 2014) da seguinte forma: cada *nó* v seleciona um limite θ_v no intervalo [0,1] uniformemente ao acaso. Assim, em cada passo de tempo t onde H_{t-1} é o conjunto de *nós* ativos no tempo $t-1$ ou anteriormente, cada *nó* inativo se torna ativo se, $\sum_{u \in \eta^{in}(v) \cap H_{t-1}} b(u, v) \geq \Theta_v$, apenas se a soma do total de pesos de seus vizinhos é pelo menos θ_v.

Na Figura 17, retirada de (SHAKARIAN, 2015), é especificada a dinâmica do

modelo. Onde a probabilidade é determinada com base nos limites fornecidos no inicio do processo (SHAKARIAN, 2015) (ZHANG, 2014), quando as arestas recebem um limite no intervalo [0,1].

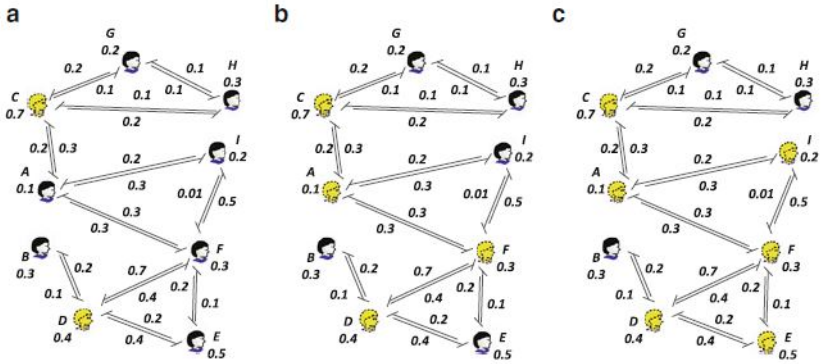


Figura 17: Dinâmica do processo de difusão no modelo *Threshold*.

Fonte: (SHAKARIAN, 2015).

Para a Figura 17a, os dois *nós*, C e D estão ativos. Assim, o *nó* C tem a possibilidade de ativar seus vizinhos G, H e A. Todavia, ele ativa apenas A $(0,2 \geq 01)$, pois seu peso para influenciar G e H não é suficiente. Ao *nó* D é dada a possibilidade de ativar os vizinhos F, B e E. Neste caso, a ativação ocorre apenas com F $(0,4 \geq 0,3)$. Na Figura 17b, quatro *nós* estão ativos e ativam I e E. No próximo passo, não há mais *nós* ativos, terminando o processo de difusão (SHAKARIAN, 2015).

4.2.2 Modelos *Independent Cascade*

Os modelos de cascata dinâmicos, embasados em estudos de sistemas de partículas interativas e teoria da probabilidade, (ZHANG, 2014), são considerados para o processo de difusão. No contexto de marketing, (GOLDENBERG; LIBAI, 2001) foram os primeiros a estudar modelos em cascata. Assim, nesses modelos, a dinâmica é capturada da seguinte forma: no instante t, quando um *nó* v se torna ativo pela primeira vez, ele tem uma chance única de influenciar cada vizinho inativo no tempo $t+1$. E caso tenha sucesso, faz com que u seja ativado com uma probabilidade $p_{v,u}$. Além disso, se vários vizinhos de u se tornarem ativos no tempo t, suas tentativas de ativar u serão processadas em uma ordem arbitrária. Se um deles, p. e., w, tiver sucesso no tempo t, então u se torna ativo no tempo $t+1$; no entanto, quer seja bem sucedido ou não, não pode fazer mais tentativas nos seguintes passos de tempo. Semelhante aos modelos *trheshold*, o processo termina quando não há mais ativações (ZHANG, 2014) (SHAKARIAN, 2015).

Independent Cascade Model em (NEWMAN, 2002) (EASLEY, 2010) (BARABASI, 2016) (ZHANG, 2014) (SHAKARIAN, 2015), em vez da probabilidade de infecção ser única para todo o processo, há uma probabilidade de infecção associada a cada aresta.

Dessa forma, $p_{u,v}$ é a probabilidade de u infectar v. Esta probabilidade pode ser atribuída com base na frequência de interações, proximidade geográfica ou traços históricos de infecção. Cada *nó*, uma vez infectado, tem a capacidade de infectar seu vizinho na próxima etapa de tempo com base na probabilidade associada a aresta.

Em (SHAKARIAN, 2015) o processo é definido da seguinte forma: a cada intervalo de tempo t onde χ_{t-1}^{new} é o conjunto de *nós* ativos iniciais no tempo $t-1$, cada $v \in \chi_{t-1}^{new}$ infecta o vizinho inativo $u \in \eta_{out} v$ com a probabilidade $p_{u,v}$.

A Figura 18 exemplifica o modelo. *Nós* ativos são exibidos em linhas pontilhadas e na cor amarela. No intervalo de tempo inicial, dois *nós* C e D são ativados. No próximo intervalo de tempo, estes *nós* têm a possibilidade de ativar três de seus vizinhos (A, G, e H) e (B, E, e F) respectivamente. De acordo com a Figura 18b, somente três *nós* A, H e E são ativados com sucesso e os *nós* ativos inicialmente mudam para a cor cinza (denotando seu estado ativo, mas sem chance de ativar outros). No próximo intervalo de tempo, dois *nós* G e F se tornam ativos e os *nós* ativos anteriores A, E e H mudam para cinza. No intervalo de tempo $t = 2$, dois *nós* F e G tornam-se ativos. Os vizinhos do *nó* G estão ativos, por isso não tem a chance de ativar nenhum *nó*. O *nó* F têm a possibilidade para ativar o *nó* I; no entanto, falha como mostrado em exemplo dado na Figura 18d. Como não há mais novo *nó* ativo, o processo de difusão parou, (SHAKARIAN, 2015).

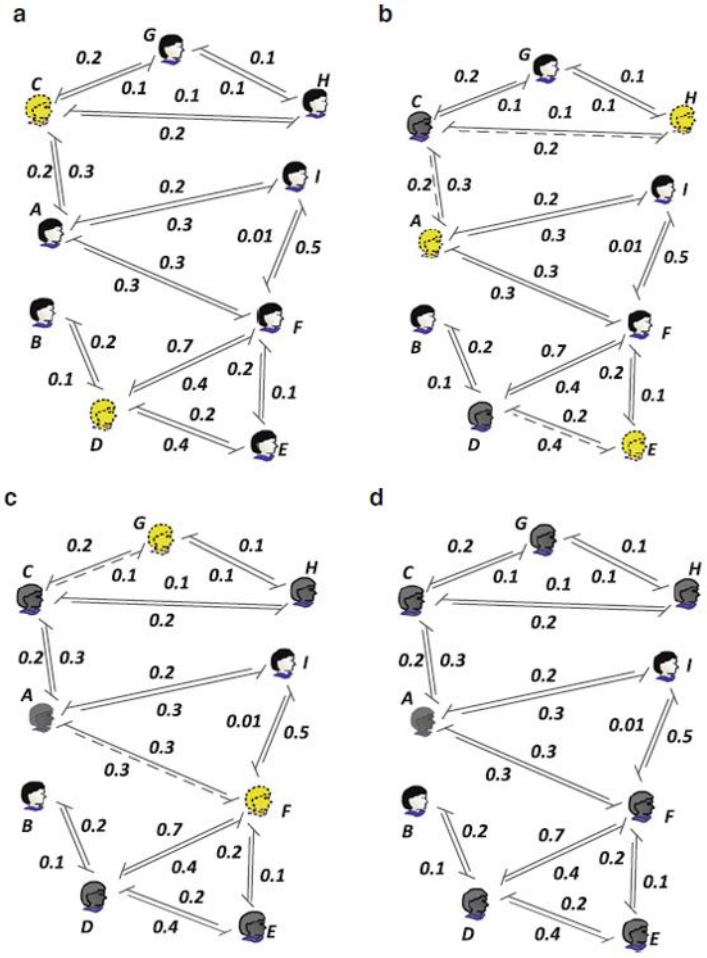


Figura 18: Dinâmica do Modelo *Independent Cascade*.

Fonte: (SHAKARIAN, 2015).

5 Metodologia

A finalidade deste capítulo é descrever o delineamento metodológico aplicado a este trabalho, visando apresentar a sequência dos passos adotados a fim de guiar a coleta e análise dos dados para a realização da pesquisa.

Por ter a intenção de comparar a taxa de difusão em redes, onde são considerados dois cenários que testam a difusão em redes diante das teorias de difusão de informação e análise de redes sociais, o estudo está categorizado de acordo com Creswell (CRESWELL; CRESWELL, 2017) pela abordagem quantitativa. Uma vez que as variáveis podem ser medidas numericamente.

A taxa de difusão considera, os dois, seguintes cenários: (i) utilizando a estrutura da rede como um todo, foram selecionados conjuntos de sementes a partir de medidas de centralidade; (ii) utilizando a estrutura de comunidades sobrepostas, ou seja, a partir da rede decomposta em comunidades, selecionou-se conjuntos de sementes segundo critérios estabelecidos para o trabalho.

A Figura 19 tem por objetivo ilustrar os passos definidos para a execução dos experimentos.

Figura 19: Ilustra os passos realizados na execução dos experimentos definidos para este trabalho.

Dessa forma, nas próximas Seções descrevem detalhadamente: a coleta de dados, o tratamento, as estratégias para seleção dos conjuntos de sementes que iniciarão o processo de difusão e caracterização do experimento.

5.1 Dados: Coleta, Tratamento e Caracterização

Esta Seção cobre os dados (redes) coletados, o tratamento dado aos mesmos que vão de encontro aos objetivos propostos e a caracterização dos mesmo de acordo com propriedades da teoria de redes.

5.1.1 Coleta de Dados

As redes selecionadas para investigação foram coletadas dos repositórios: SNAP (LESKOVEC; KREVL, 2014), Projeto KONECT (KUNEGIS, 2013), Network Data (ROBBIN, 2015) e Network Repository (ROSSI; AHMED, 2015), para as quais segue uma pequena descrição.

Do repositório SNAP foram selecionados os seguintes conjuntos de dados:

Wiki-Vote - é uma rede de colaboração, de uma enciclopédia gratuita Wikipedia, escrita por voluntários em todo o mundo. Uma pequena parte dos colaboradores desta enciclopédia são administradores, cujo papel possibilita acesso a recursos técnicos adicionais para manutenção da mesma. Para que um usuário se torne administrador, é emitida uma Solicitação de administração (RfA) e a comunidade da Wikipedia, por meio de uma discussão pública ou de uma votação, decide quem deve ser promovido à administração. Os *nós* da rede representam os usuários da Wikipedia e as arestas direcionadas do *nó* i para o *nó* j representa que o usuário i votou no usuário j.

CA-GrQc - rede de colaboração cientifica do arXiv entre autores de artigos digitais da seção Relatividade Geral e Cosmologia Quântica. Se um autor i é coautor de um artigo com o autor j, o grafo contém uma aresta não direcionada de i a j.

Soc-Epinions1 - é uma rede social *on-line*, de confiança na avaliação geral de consumidores hospedada no site Epinions.com. A rede é direcionada, pois os membros do site podem decidir se 'confiam' uns nos outros. Todas as relações de confiança interagem e formam a *Web of Trust*, que é então combinada com avaliações de avaliações para determinar quais avaliações são mostradas ao usuário.

Soc-sign-epinions - como a soc-Epinions1, é uma rede online, direcionada, de confiança na avaliação geral de consumidores, no *site* Epinons.com. Membros do site podem decidir se "confiam" uns nos outros. Entretanto, esta rede é assinada, ou seja, os membros declaram se confiam ou não na opinião dos outros.

Ca-HepTh - é uma rede da seção Teoria da Física de Alta Energia. A seção cobre colaborações científicas entre artigos de autores submetidos a essa categoria. Se um autor i é coautor de um artigo com o autor j, o grafo contém uma aresta não direcionada de i a j.

Do projeto KONECT apenas a seguinte foi utilizada:

Moreno_innovation - é uma rede direcionada que captura a inovação difundida entre 246 médicos em cidades do estado de Illinois, em USA. Cada *nó* da rede representa um médico e a aresta é definida entre dois médicos de duas formas: se i é amigo de j, ou i se aconselha ou esta interessado em uma discussão com j. Mesmo que ambas as condições forem satisfeitas, existe apenas uma aresta entre dois médicos. A rede mantém um relacionamento de confiança.

Do repositório Network Data (ROBBIN, 2015) foram coletados os seguintes conjuntos de dados:

Karate - rede social de amizade, não direcionada, entre 34 membros de um clube de karate de uma universidade dos EUA.

Dolfins - rede social, não direcionada, da associação entre 62 golfinhos habitando na comunidade de Doubtful Sound, Nova Zelândia.

Polbooks - rede de livros sobre política dos EUA. Os nós representam livros sobre a política dos Estados Unidos vendidos pela livraria online Amazon.com. As arestas representam a frequência da compra de livros pelos mesmo compradores, conforme indicado pelos "clientes que compraram este livro também compraram esses outros livros"aparecem na Amazon.

Por fim, do Network Repository foram utilizadas as redes:

Football - dados que contém a rede, não direcionada, de jogos de futebol americano entre faculdades da Divisão IA durante a temporada regular do outono 2000.

Jazz - é uma rede de músicos de Jazz. Esta catalogada dentro no repositório em redes diversas. A mesma é uma rede não direcionada.

5.1.2 Extração da componente gigante

Das características estruturais de uma rede, uma de particular importância é a noção de componentes conectados, pois eles indicam o quão forte e sustentáveis as interações são estabelecidas (AYDIN; PERDAHCI, 2016). Assim, o maior desses componentes, definido na literatura, como componente gigante[1], é extraído dos dados coletados pois o mesmo desempenha um papel fundamental para o estudo de difusão deste trabalho. Uma vez que é necessário garantir que nenhuma semente selecionada para iniciar o processo não encontre restrição quanto ao alcance de um componente desconectado na rede.

[1] Componente gigante é o maior componente conectado da rede considerando o número de *nós* e está relacionado ao modelo de Erdos Reniy, para maiores detalhes veja (ERDŐS; RÉNYI, 1960).

5.1.3 Normalização dos dados

A normalização dos dados nada mais é do que definir uma sequência de inteiros para os *nós* das redes. Ou seja, retorna uma cópia da rede com os *nós* rotulados usando inteiros consecutivos. Alguns dos conjuntos de dados coletados, tais como: moreno_innovation, soc-Epinions1, Wiki-Vote e soc-sign-epinions também passaram pelo processo de inversão de arestas antes da normalização. Isto se deve ao fato de que são redes de confiança (um indivíduo vota em quem confia) direcionadas. Dessa forma, o indivíduo que "votou"foi influenciado por aquele que recebeu o "voto", o que torna necessária a inversão de arestas para o proposto neste trabalho.

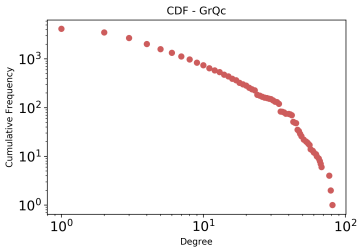
5.1.4 Caracterização dos dados

A Tabela 5.1.4 contém descrições das informações básicas sobre as redes reais. A tabela contém para cada rede o seu identificador (REDE), o número de *nós* (N), número de arestas (M), a densidade da rede (DENS), a média dos graus de entrada e saída ou simplesmente grau médio (GRAU MÉDIO) e o coeficiente de *clustering* (C). As estatísticas apresentadas para cada rede, considera a componente gigante, como nos trabalhos de (SHAKARIAN, 2015) e (SALAVATI, 2018), pois a falha na cascata vai ocorrer se não existir sementes em todas as componentes.

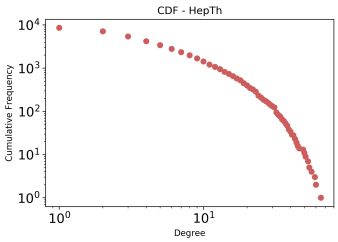
	N	M	GRAU MÉDIO	DENS.	(C)
CA-GrQc	4158	13428	6.46	0.0015	0.556
CA-HepTh	8638	24806	5.74	0.0006	0.481
Dolphins	62	159	5.13	0.084	0.258
Football	35	118	6.74	0.198	0.338
karate	34	78	4.588	0.139	0.570
Jazz	198	2742	27.697	0.140	0.617
Polbooks	62	159	5.129	0.084	0.258
Moreno_Innovation	117	465	3.97/3.97	0.034	0.109
Epinions1	75877	405739	5.35/5.35	0.000070	0.069
Sign-epinions	119130	704572	5.19/5.19	0.00005	0.070
Wiki-Vote	7066	100736	14.25/14.25	0.0020	0.071

Tabela 1: Informações Básicas das Redes Reais

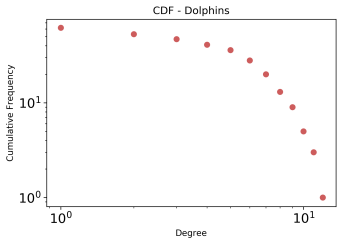
A Figura 20, traz os gráficos da função de distribuição acumulada de grau para cada uma das redes reais de estudo. Os gráficos mostram que as redes possuem distribuição *power-law*. De acordo com Lópes (LÓPEZ-PINTADO, 2008), isso implica que cada *nó* tem uma probabilidade estatisticamente significativa de possuir um número muito grande de conexões em relação a conectividade média da rede, o que torna a conectividade da rede heterogênea.



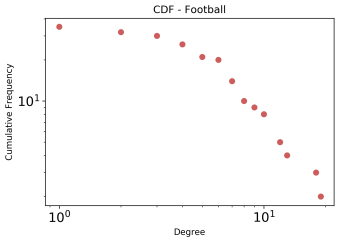
(a) CDF - CA-GrQc.



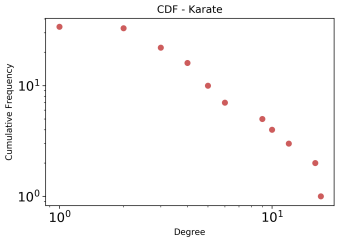
(b) CDF - CA-HepTh.



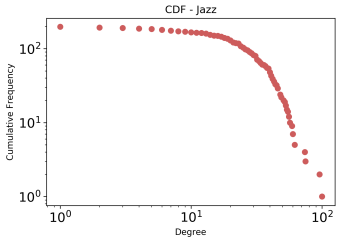
(c) CDF - Dolphins.



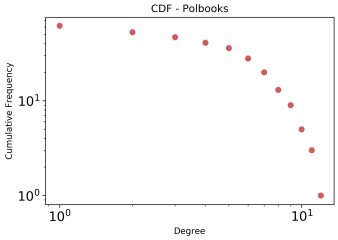
(d) CDF - Football.



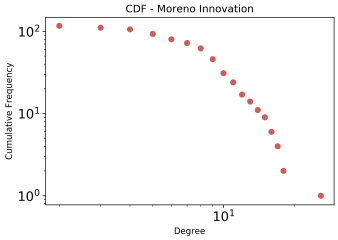
(e) CDF - Karate.



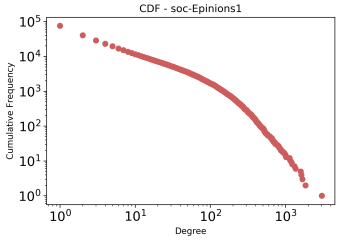
(f) CDF - Jazz.



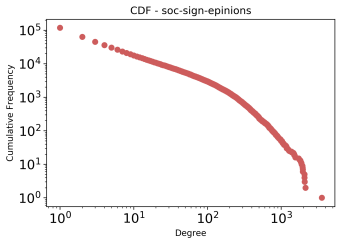
(g) CDF - Polbooks.



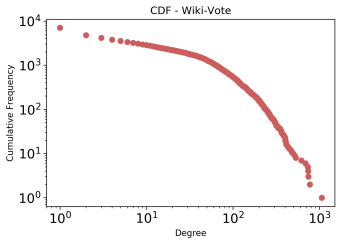
(h) CDF - Moreno.



(i) CDF - soc-Epinions1.



(j) CDF - soc-sign-epinions



(k) CDF - Wiki-Vote.

Figura 20: CDF das Redes Reais.

Redes sintéticas também foram utilizadas na comparação de medidas de centralidade e critérios utilizando comunidades sobrepostas. As redes sintéticas de Lancichinetti *et al* (LANCICHINETTI, 2008) seguem uma série de propriedades que se parecem com as redes reais e por isso o *benchmark* LFR é tão utilizado na literatura, o que definiu a escolha deste *banchmark* para este trabalho.

As configurações para a geração das redes seguem parâmetros similares aos apresentados pelos autores. Dessa forma, os parâmetros utilizados para a geração das redes foram: (n) para o número dos *nós*; (τ_1) define o expoente, *power-law* definido na literatura como valores entre 2 e 3 (encontrados em redes reais), que representa o grau de distribuição da redes; τ_2 também define o grau de distribuição, *power-law*, mas neste caso, para as

comunidades, cujos valores defendidos pelos autores se encontra nos intervalos de 1 e 2; o valor de μ_u possui o proposito delinear a fração de arestas incidentes intra-comunidades, cujo valor deve estar entre os intervalos [0,1]; o grau mínimo e o grau máximo para os *nós*, o número mínimo e máximo de comunidades e o número máximo de iterações para que o algoritmo alcance o tamanho das comunidades, grau de distribuição e afiliações. Na tabela 2 são exibidas as configurações para a geração das redes, neste caso, as redes foram nomeados com o acrônimo formado pelo nome dos autores (LFR), seguido por identificador numérico obtidos da sequência das execuções. O total de redes geradas foram 16, divididas em dois blocos, modificando apenas o parâmetro μ_u, o primeiro bloco com o valor de 0.1 e o segundo com 0.2.

	LFR1	LFR2	LFR3	LFR4	LFR5	LFR6	LFR7	LFR8
n	1000	1000	3000	3000	1000	1000	3000	3000
τ_1	2.0	2.5	2.0	2.5	2.0	2.5	2.0	2.5
τ_1	1.2	1.5	1.2	1.5	1.2	1.5	1.2	1.5
μ_u	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
min_degree	5	5	5	5	5	5	5	5
grau máximo	50	50	50	50	50	50	50	50
min_coms	n*0.25	n*0.25	n*0.25	n*0.25	n*0.25	n*0.25	n*0.25	n*0.25
max_com	n*0.25	n*0.25	n*0.25	n*0.25	n*0.25	n*0.25	n*0.25	n*0.25
n_iterações	1000	1000	1000	1000	1000	1000	1000	1000

Tabela 2: Configurações para geração de redes sintéticas de acordo com o modelo Lanchicinetti et al.

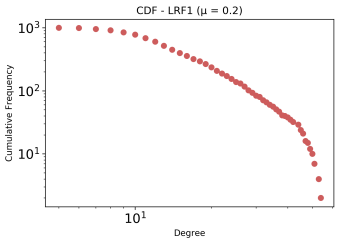
Nas Tabelas 3 e 4, estão dispostas as informações básicas para tais redes. A apresentação das duas tabelas é devido as redes resultantes com variação do parâmetro μ_u. Assim, respectivamente estão apresentadas as informações básicas para $\mu_u = 0.1$ e $\mu_u = 0.2$.

Dessa forma passa-se para o segundo passo, que é identificar as sementes para os conjuntos de redes.

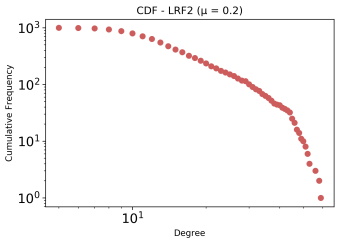
	N	M	GRAU MÉDIO	DENS.	(C)
LFR1	1000	7489	14.97	0.015	0.059
LFR2	1000	7872	15.74	0.016	0.068
LFR3	1000	6161	12.32	0.012	0.046
LFR4	1000	6261	12.52	0.012	0.051
LFR5	3000	23807	15.87	0.005	0.022
LFR6	3000	23697	15.79	0.005	0.023
LFR7	3000	19017	12.67	0.004	0.016
LFR8	3000	19045	12.69	0.004	0.016

Tabela 3: Informações Básicas das Redes LFR com $\mu_u = 0.1$

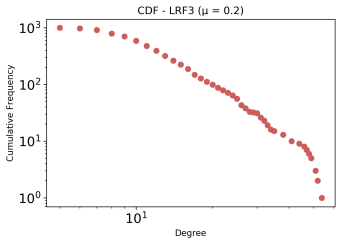
É exibido nas Figuras 21 e 22 os gráficos da função de distribuição acumulada de grau para cada uma das redes sintéticas (LFR), confirmando a propriedade de *power-law*.



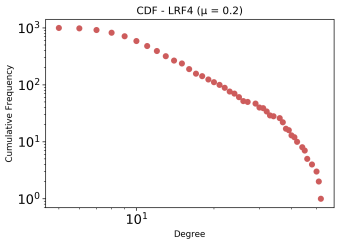
(a) CDF - LFR1.



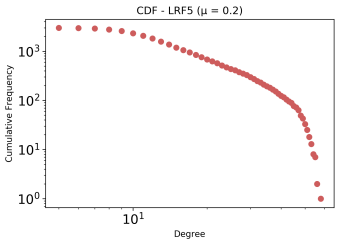
(b) CDF - LFR2.



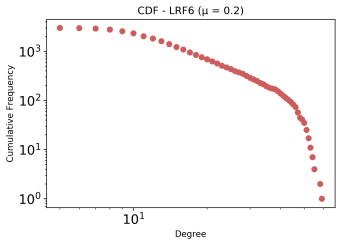
(c) CDF - LFR3.



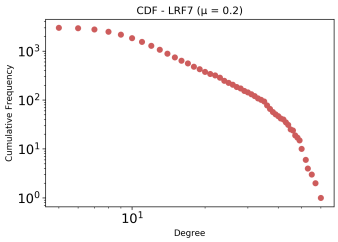
(d) CDF - LFR4.



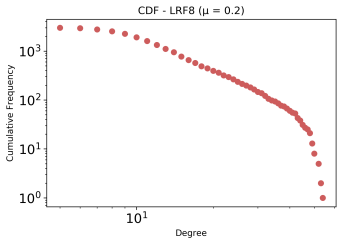
(e) CDF - LFR5.



(f) CDF - LFR6.



(g) CDF - LFR7.



(h) CDF - LFR8.

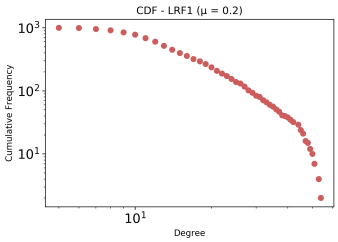
Figura 21: CDF das Redes Sintéticas para $\mu_u = 0.1$.

	N	M	GRAU MÉDIO	DENS.	(C)
LFR1	1000	7887	15.77	0.015	0.045
LFR2	1000	8104	16.20	0.016	0.048
LFR3	1000	6041	12.08	0.012	0.0319
LFR4	1000	6193	12.38	0.012	0.032
LFR5	3000	23987	15.99	0.005	0.017
LFR6	3000	24032	16.02	0.005	0.016
LFR7	3000	19314	12.87	0.004	0.012
LFR8	3000	19690	13.12	0.004	0.011

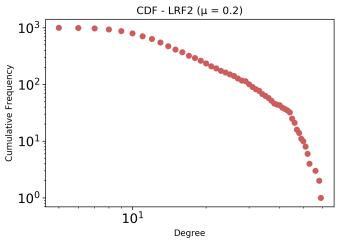
Tabela 4: Informações Básicas das Redes LFR com $\mu_u = 0.2$

5.2 Identificação do conjunto de sementes

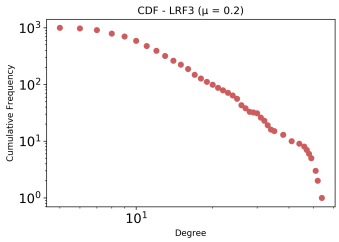
Como extrair um pequeno conjunto de *nós* (ou sementes) de uma rede que podem gerar máxima propagação sob um modelo de difusão? Na literatura não foi encontrado uma definição da quantidade mínima ou máxima de elementos para compor o conjunto de sementes. Dessa forma, para este trabalho foi definido que o conjunto de sementes teria 10% dos elementos da rede, definido como *qtd_seeds*.



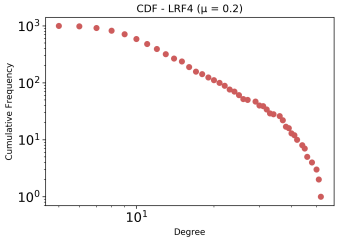
(a) CDF - LFR1.



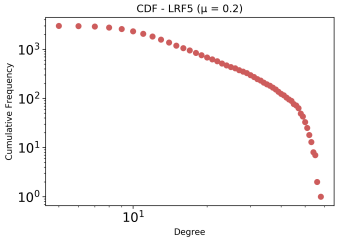
(b) CDF - LFR2.



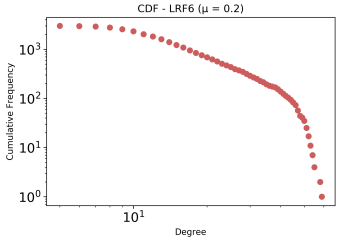
(c) CDF - LFR3.



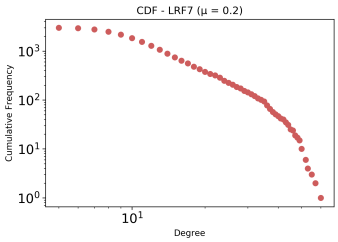
(d) CDF - LFR4.



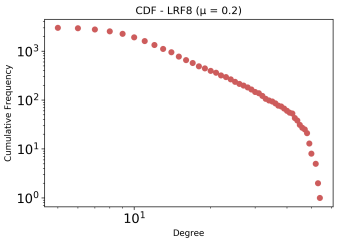
(e) CDF - LFR5.



(f) CDF - LFR6.



(g) CDF - LFR7.



(h) CDF - LFR8.

Figura 22: CDF das Redes Sintéticas para $\mu_u = 0,2$.

5.2.1 Identificando sementes de acordo com medidas de centralidade

Na busca por *nós*, nas redes, candidatos a sementes para iniciarem um processo de difusão foram utilizas algumas medidas de centralidade. Uma vez que as mesmas tem sido muito utilizadas na literatura para classificar a habilidade de propagação dos *nós* de acordo com sua posição topológica (SALAVATI, 2018). Assim, foram usadas as medidas de centralidade de grau, *betweenness*, *closeness*, *PageRank* além de sementes selecionadas aleatoriamente.

5.2.2 Identificando sementes de acordo com informações de comunidades sobrepostas

Para identificar sementes a partir da estrutura de comunidade das redes, foram executados os algoritmos de detecção de comunidades SLPA (XIE, 2011) e DEMON (COSCIA, 2012). A escolha se deu por ambos possuírem como parâmetro obrigatório apenas a rede. Dos resultados obtido destas execuções (conjuntos de comunidades), três critérios para a seleção de sementes foi estabelecido para cada método.

A execução dos algoritmos SLPA e DEMON se deram por meio de suas implementações disponibilizadas na biblioteca CDlib (*Community Discovery Library*), versão 0.1.7) (ROSSETTI, 2019). A biblioteca possui uma gama de funções *fitness* que permitem resumir características de um conjunto de comunidades resultantes da execução. Dessa forma, o método *pool_grid_filter* foi selecionado porque executa internamente os métodos quantas vezes forem determinadas por intervalos de valores. No caso dos métodos utilizados, intervalos de parâmetros opcionais. Retornando a partição ideal para cada algoritmo de acordo com uma função de qualidade especificada. Para este trabalho foi escolhida a medida de modularidade de (NICOSIA, 2009) implementada pela biblioteca. Para o DEMON, $\epsilon\%$ (parâmetro que controla o tamanho das sobreposições) a biblioteca define 0.25 como padrão. Assim , para este método foi especificado o intervalo de de ϵ em $\{0,25, ..., 0,45\}$, incrementando o valor inicial $0,25$ em 1. O que gerou 3 execuções. Já para o método SLPA, o parâmetro t (critério de parada com um número máximo de iterações) foi especificado no intervalo de 10 a 22, ou seja, $\{10, 11, 12, ..., 22\}$, uma vez que os autores afirmam que para redes grandes, 21 iterações é o suficiente. O parâmetro r que controla os rótulos armazenados pelos *nós* foi especificado para os intervalos $\{0,1, 0,2, ..., 0,5\}$.

Tirando proveito das execuções para descoberta de comunidades, também foram extraídas as seguintes informações: medida OMNI (*ovelapping mutual normalized information*)[2] de (LANCICHINETTI, 2009) e gráficos que exibem a distribuição da densidade interna de arestas nas comunidades, esta última referente a cada método. Informações sobre os resultados do processo de detecção de comunidades podem ser encontrados no Anexo A.

Após a obtenção das comunidades sobrepostas, sementes foram selecionadas de acordo com os seguintes critérios:

1. Critério A - para cada comunidade c_i obtida do conjunto de comunidades como resultado da execução dos métodos, foi calculado um valor, resultante da razão do tamanho de cada c_i em relação ao tamanho da rede. Assim, $p_i = \frac{tamanho_c_i}{tamanho_rede}$. Posteriormente, utiliza-se esta razão para determinar o número de sementes com o qual cada comunidade c_i irá contribuir para iniciar o processo de difusão, que é $p_i * qtd_seeds$. Nos casos onde o produto de $pi * qtd_seeds$ é menor que 1, faz com que uma específica comunidade c_i não contribua com nenhum elemento. Para contornar o problema possibilitando que os 10% de sementes fossem selecionadas, cada comunidade cujo valor de contribuição é menor que 1, automaticamente passa a contribuir com 1 elemento. Este critério considera a estrutura de comunidade com sobreposição, mas não os elementos que estão na sobreposição.

2. Critério B - para este critério os seguintes passos foram executados: (1) para cada

[2] ONMI- medida utilizada para verificar a similaridade do resultado das comunidades encontradas entre os métodos.

conjunto de comunidades é realizada uma combinação de c_i, ou seja, c_i1 com c_i2, até c_in, c_i2 com c_i3 até c_in e assim por diante. Para cada combinação é extraído conjuntos de *nós* comuns às comunidades, obtendo aqueles que estão nas sobreposições; (2) a partir dos conjuntos sobrepostos é utilizado a estrutura de dicionário (da linguagem python) para armazenar um conjunto de *nós* onde é descrito o elemento (*nó*) com o respectivo valor que determina o número de ocorrências com a qual o mesmo aparece nas sobreposições; (3) o conjunto resultante do passo 2 foi ordenado tendo como chave o valor obtido do número de ocorrências em ordem decrescente. Dessa forma os *nós* que participam de mais comunidades são selecionados para compor o conjunto de sementes para este critério.

3. Critério C - segue o princípio do critério B, selecionar *nós* que estão nas sobreposições. Embora, neste caso, ao conjunto resultante das sobreposições é atribuído a cada *nós* o seu grau na rede. É realizada a ordenação do conjunto de *nós* em ordem decrescente de acordo com o grau mais alto. Dessa forma, têm-se para o conjunto de sementes, elementos encontrados nas sobreposições que possuem maior grau.

5.3 Descrição dos Experimentos

Para cada critério e métrica executados para a seleção de sementes mencionados acima, buscou-se escolher os s conjunto de melhores sementes (10%) do total de elementos presentes em cada rede, de acordo com a natureza de cada um. Para os conjunto s encontrados, obtém-se 5 subconjuntos, contendo os primeiros 20%, 40%, 60%, 80% e 100%. Por fim, cada subconjunto é submetido ao processo de difusão utilizando os modelos *Independent Cascade* e *Threshold*. Dessa forma, para cada rede obtém-se um total de 11 execuções para cada subconjunto. Ou seja, as quatro métricas de centralidade, a seleção por aleatoriedade e os três critérios propostos executados para cada algoritmo de detecção de comunidades, SLPA e Demon.

O processo de difusão foi realizado para cada subconjunto utilizando a implementação dos modelo citados, disponíveis na biblioteca NDlib (ROSSETTI, 2018) versão 5.0.0. Os processos de difusão foram executados variando o parâmetro de limite para cada aresta de $0,1$ até $0,6$. Os exemplos disponibilizados na biblioteca exibem $0,1$ para o *Independent Cascade* e $0,25$ para o *Threshold*.

Cada experimento, utilizando um subconjunto resultou em uma porcentagem do total de *nós* influenciados na rede. Dessa forma, a avaliação foi feita comparando os resultados obtidos das métricas com os resultados dos critérios em ambos os modelos de difusão (*Independent Cascade* e *Linear Threshold*).

Os experimentos foram conduzidos em ambiente Linux, processador Intel (R) Core (TM) i7 de 64 bits, 2.80 GHz, equipado com 16 GB de RAM. O código foi desenvolvido

em Python 3.6. As bibliotecas, NetworkX (versão 2.3) e Igraph (versão 0.7.1), ambas em Python destinadas a criar, manipular e estudar a estrutura e a dinâmica de redes complexas. A biblioteca Cdlib (*Community Detection Library*) (versão 0.1.7) para extrair, comparar e avaliar comunidades de redes complexas. Já a biblioteca NDlib (*Network Diffusion Library*) (versão 5.0.0) para descrever, simular e estudar processos de difusão em redes complexas. Por fim o ambiente Jupyter Notebook que une texto e código, foi utilizado para extrair informações estatísticas das redes e gerar gráficos (visualizações) de informações obtidas no trabalho.

6 Resultados e Discussões

A comparação do efeito entre métricas e critérios no processo de difusão é avaliada com base no percentual de propagação alcançado por cada um. Ou seja, quanto maior a propagação, melhor é a qualidade do conjunto de semente. Este tipo de avaliação está presente na literatura (KEMPE, 2003) (GOYAL, 2011) (CHEN, 2009) (LIN, 2015) (MILLI, 2018) (ZHANG, 2014). Também foram criadas relações entre as sementes e arestas (criadas a partir dos mesmos gerando um subgrafo). Dessa forma, além do percentual de difusão, é possível verificar as relações entre as sementes selecionadas de acordo com métricas e critérios propostos.

Os resultados dos experimentos geraram 324 tabelas, cada uma delas contendo 55 resultados das execuções das métricas (5) e critérios (6) para cada subconjunto de sementes selecionados (veja Seção 5.3). Para simplificar a análise optou-se em apresentar apenas os resultados para as redes, quando o valor de limite para os modelos está de acordo ou próximo ao utilizado na biblioteca NDlib. Dessa forma são utilizados os resultados para o modelo ICM na configuração de $0,1$ e o modelo LTM na configuração de 0.3 (próximo aos $0,25$) apresentados na biblioteca. Resultados para as redes sintéticas LFR com $\mu_u = 0,2$ foram descartados, pois o aumento deste parâmetro criou redes onde as comunidades não são bem definidas, o que em alguns casos resultou em apenas uma comunidade para toda a rede.

Os resultados das execuções da difusão, sob os modelos propostos, para cada rede, quando todos os elementos (100%) dos conjuntos de sementes são utilizados, foi normalizado de acordo com o valor da métrica e/ou critério com melhor desempenho no intervalo [0,1]. Ou seja, o(s) resultado(s), cujas métricas e/ou critérios obtiveram o melhor resultado receberam o valor de 1. A seguir serão apresentados os resultados para os modelos *Independent Cascade* e *Linear Threshold*, respectivamente.

6.1 Resultados da Difusão sob o Modelo ICM

Nesta Seção, para o modelo *Independent Cascade* são disponibilizados os resultados para as redes reais, na tabela 6.1.1 e sintéticas na tabela 6.1.2. Na última linha das referidas tabelas estão também contabilizados os somatórios para métricas e critérios em todas as redes sob o índice "Total".

6.1.1 Redes Reais

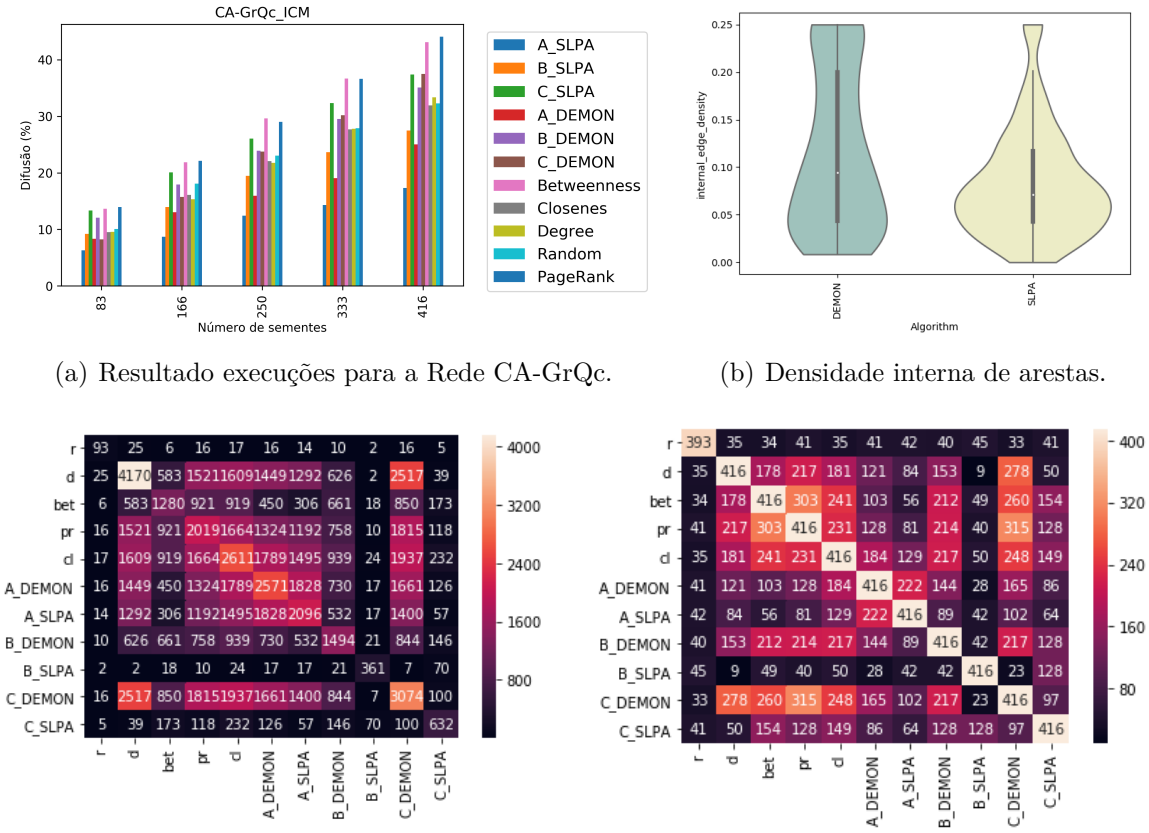
	A _SLPA	B_SLPA	C_SLPA	A_DEMON	B_DEMON	C_DEMON	BC	CC	DC	Random	PageRank
CA-GrQc	0.39	0.62	0.85	0.57	0.80	0.85	0.98	0.72	0.76	0.73	1.00
CA-HepTh	0.54	0.65	0.80	0.67	0.79	0.91	0.96	0.76	0.88	0.73	1.00
Dolphins	0.83	0.00	0.00	0.65	0.65	0.91	0.78	0.70	1.00	0.83	0.83
Football	0.57	0.00	0.00	0.86	0.00	0.00	0.76	1.00	1.00	0.48	0.86
Jazz	0.94	0.00	0.00	0.91	0.97	0.98	1.00	0.93	0.97	0.97	0.98
karate	0.94	0.00	0.00	0.81	0.94	0.44	0.81	0.94	1.00	0.62	0.87
Moreno											
Innovation	0.53	0.35	0.61	0.51	0.88	0.84	0.76	0.92	1.00	0.43	0.29
Polbooks	0.58	0.33	0.46	0.50	0.42	0.71	0.62	0.54	0.83	0.75	1.00
Epinions1	0.94	0.67	0.75	0.94	0.93	0.98	1.00	0.96	0.99	0.68	0.70
Sign-epinions	0.96	0.53	0.59	0.97	0.96	1.00	0.99	0.98	1.00	0.66	0.72
Wiki-Vote	0.92	0.72	0.89	0.93	0.69	0.99	0.96	0.99	1.00	0.76	0.59
Total	8.14	3.87	4.95	8.32	8.03	8.61	9.62	9.44	10.43	7.64	8.84

Tabela 5: Resultados para as redes reais com execução sob o Modelo *Independente Cascade Model*, cujo valor do limite (*threshold=0,1* é o utilizado na biblioteca NDlib. As letras A, B e C antes do nome dos métodos de detecção de comunidades (SLPA e DEMON) se referem aos critérios propostos para o trabalho de acordo com as execuções para cada método. Onde se lê BC, é *Betweenness Centrality*; CC, *Closeness Centrality* e DC, *Degree Centrality*.

Como pode ser observado na Tabela 6.1.1, algumas células apresentam valores zerados. Isso ocorreu quando o critério não obteve a contribuição de sementes esperadas. Este fato ocorreu em redes onde os algoritmos de detecção de comunidades ao executar sobre a rede, retornou apenas uma comunidade. Ou seja, a rede é uma única comunidade e, consequentemente sobreposições não existem. Em (LANCICHINETTI; FORTUNATO, 2009) um bom algoritmo de detecção de comunidades deve permitir que se encontre não apenas comunidades, se as mesmas existem, mas também 'pseudo-comunidades', quando a rede não possui estrutura de comunidade real. Neste caso, para a rede Football, os dois métodos de detecção falharam em encontrar comunidades com sobreposição, bem como o método SLPA para as redes, Dolphins, Jazz e Karate.

As redes CA-GrQc e CA-Hepth obtiveram comportamento similar no processo de difusão para o modelo em questão. Portanto para minimizar repetições de resultados, escolhemos a rede CA-GrQc para detalha-los, ilustrados na Figura 23. A Figura 23(a) responde pelos resultados das execuções no processo de difusão para todos os conjuntos de sementes, no eixo x estão descritos os subconjuntos das sementes, bem como a cor definida para cada medida de centralidade ou critério aplicados. O eixo y representa o percentual de elementos influenciados para cada subconjunto na de acordo com o tamanho rede. A Figura 23(b)mostra a avaliação da densidade interna de arestas nas comunidades resultantes da execução dos métodos SLPA e DEMON. O gráfico (*violin plot*), permite visualizar a distribuição e a densidade de probabilidade dos dados, ou seja, das comunidades. Já as Figuras 23(c) e 23(d) mostram as relações entre os subconjuntos de arestas, obtidos de um subgrafo gerado a partir de cada subconjunto de sementes e as relações entre os subconjuntos de sementes, respectivamente.

Em conjunto com a Tabela 6.1.1 e a Figura 23(a) é fácil identificar que os melhores resultados obtidos no processo são identificados para as métricas *PageRank* e *Betweenness*, seguidas pelos critérios C_Demon, C_Slpa e B_Demon em detrimento a outras métricas. Na diagonal principal da matriz na Figura 23(c), é mostrado a quantidade de arestas do subgrafo gerado a partir de cada subconjunto de arestas. A paleta de cores indica o nível com o qual cada subconjunto intersecciona com os outros, quanto mais claro, maior é o número de arestas em comum entre eles. A partir desta observação, vê-se que o número de arestas para o subconjunto de grau é maior que para os outros, seguido do subconjunto C_Demon. Entretanto, como apresentado na literatura, nem sempre o mais conectado é o mais importante na rede. Apesar de 55% dos subconjuntos de arestas das medidas de centralidade *PageRank* e *Betweenness*, esta última apresenta desempenho pouco menor que o desempenho da medida *PageRank*, o que torna evidente que a medida de *Betweenness* possui sementes importantes para o processo de difusão. Fato que pode ser observado pelo melhor desempenho deste subconjunto na terceira execução do processo de difusão. Um ponto a ser considerado em relação ao desempenho inferior dos critérios, pode estar relacionado a distribuição das comunidades em relação a densidade de arestas. No



(a) Resultado execuções para a Rede CA-GrQc.

(b) Densidade interna de arestas.

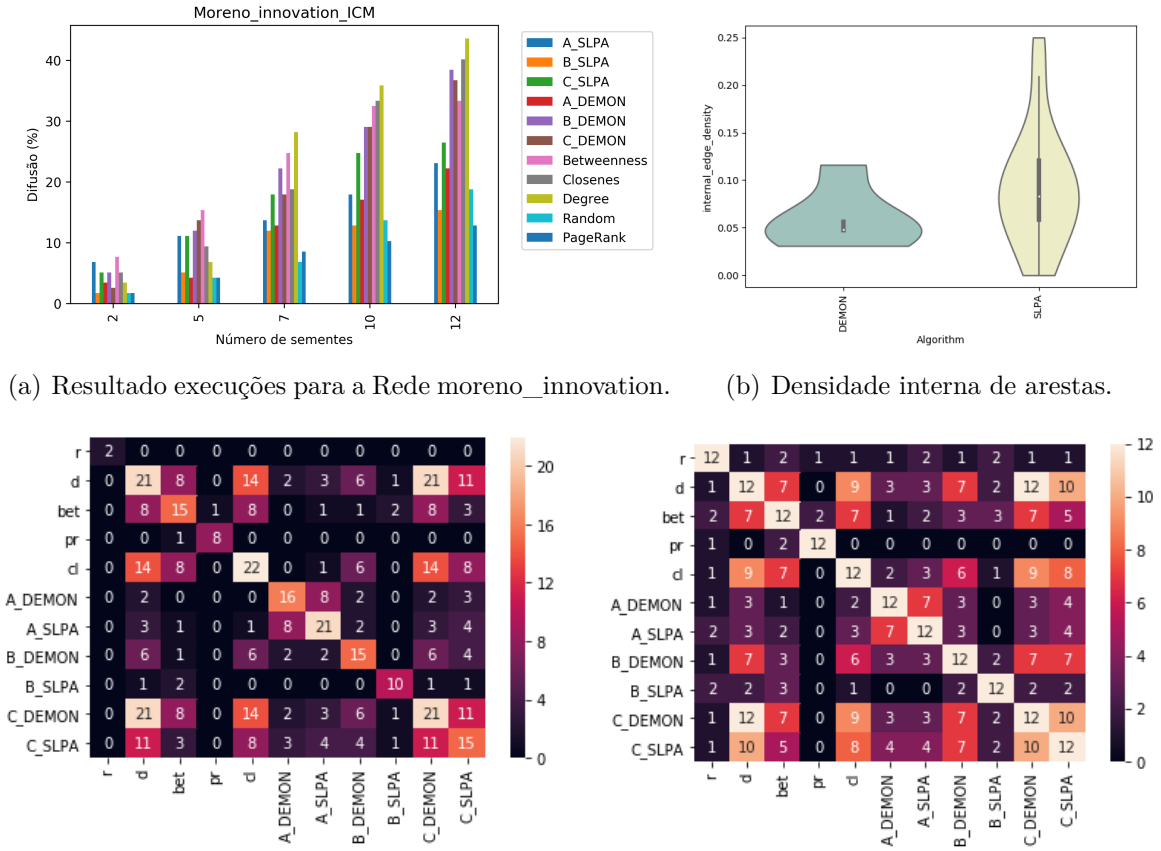
(c) Relação de Arestas.

(d) Relação de Sementes.

Figura 23: Resultados alcançados para a rede GA-GrQc sob o modelo ICM.

caso da rede CA-GrQc, mostra que a densidade interna média das arestas é menor para o SLPA que para o DEMON. A forma da distribuição indica que no caso das comunidades obtidas com o SLPA, há poucas comunidades com alta densidade interna de arestas (em torno de 0,20), diferente de boa parte das comunidades cuja distribuição se concentra em trono da mediana (ponto branco no meio da barra mais escura no centro da forma. Já a Figura 23(d), a relação entre os subconjuntos de sementes, permite observar pela coloração das células da matriz, o grau de similaridade entre os conjuntos de sementes. Assim, entre o melhor subconjunto (*PageRank*) e *Betweenness* existem 303 sementes em comum, já para C_Demon, são 315. A diferença de doze sementes entre o segundo e terceiro classificado, provavelmente se deve a sementes do subconjunto *Betweennes* serem pontes entre as comunidades. Já a diferença entre o primeiro e quarto classificados é bem maior, o C_Slpa possui 128 sementes em comum com o *PageRank*.

Na figura 24 tem-se um resumo que auxilia na análise dos resultados para a rede moreno_innovation. A Figura 24(a), mostra a classificação dos melhores resultados na difusão, respectivamente, grau,*closeness*, B_Demon e C_Demon, considerando que o alcance máximo de difusão atingiu pouco mais de 40% da rede. No contexto de *marketing*, garantir a difusão de um produto utilizando a métrica de grau, exige que tanto as sementes quanto seus vizinhos diretos recebam o produto, o que seria custoso, visto que há 22



(a) Resultado execuções para a Rede moreno_innovation.

(b) Densidade interna de arestas.

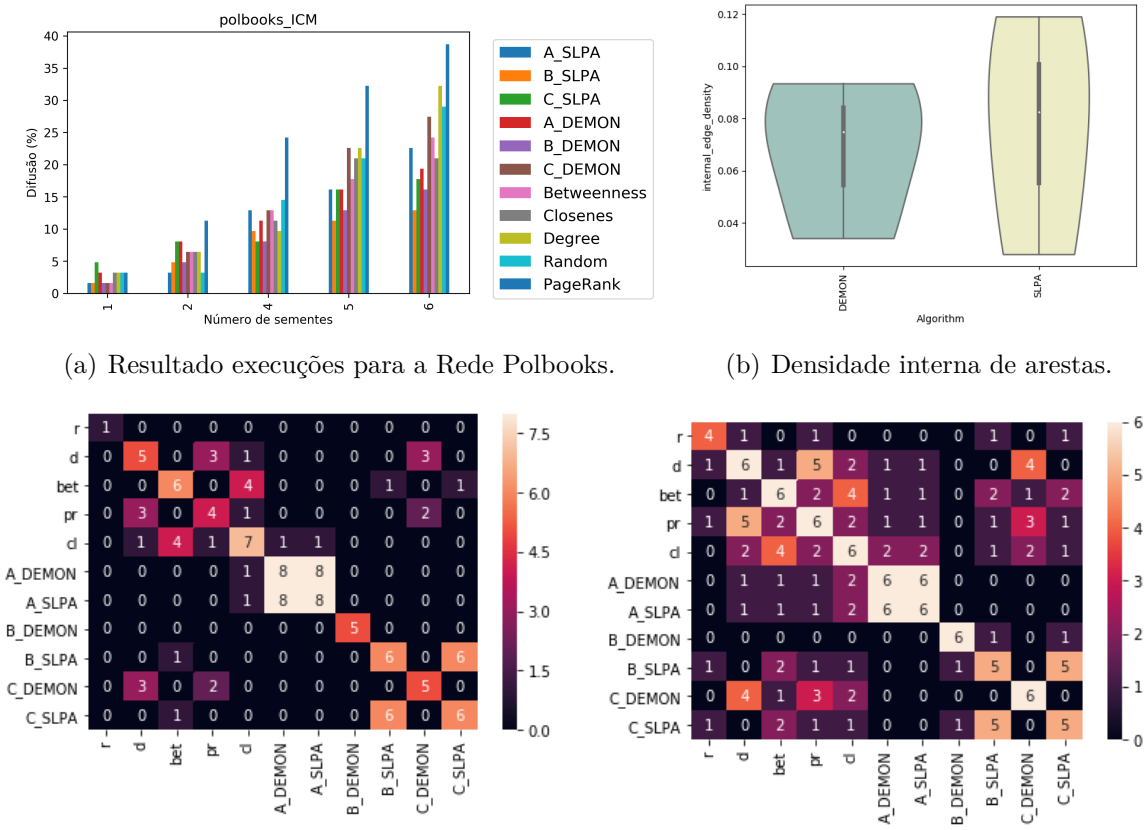
(c) Relação de Arestas.

(d) Relação de Sementes.

Figura 24: Resultados alcançados para a rede moreno_innovation sob o modelo ICM.

arestas no subgrafo gerado a partir das 12 sementes. Dos outros classificados, B_Demon é o que oferece a menor quantidade de arestas, fazendo com que seja uma possível solução, para este contexto, caso se deseje minimizar o custo mesmo que o alcance seja menor. Apesar da métrica de grau e critério C_Demon compartilharem os mesmos números de arestas e sementes, porém o último com menor desempenho, provavelmente se deve a reorganização das arestas pelo algoritmo de detecção de comunidade. De maneira geral, as sementes que não estão na interseção dos outros subconjuntos em relação ao subconjunto do grau, são aquelas que melhoram a difusão.

A rede Polbooks, Figura 25(a), na execução com os 100% do conjunto de sementes, como as redes CA-GrQC e HepTh, também obtém da métrica *PageRank* o melhor resultado no processo de difusão seguido pelo subconjunto de grau e de escolha aleatória. A eficiência do subconjunto de escolhas aleatórias se deve ao tamanho da rede. Assim, para 20% das sementes selecionadas, neste caso, pode-se afirmar que o critério C_Slpa supera os demais subconjuntos considerando que o mesmo possui 1 elemento a menos. O que devido ao tamanho da rede, pode fazer grande diferença, uma vez que com um único elemento, este critério supera em aproximadamente 65% a métrica de grau para a primeira execução. Outra observação vai para o fato de que o melhor resultado para o critério do



(a) Resultado execuções para a Rede Polbooks.

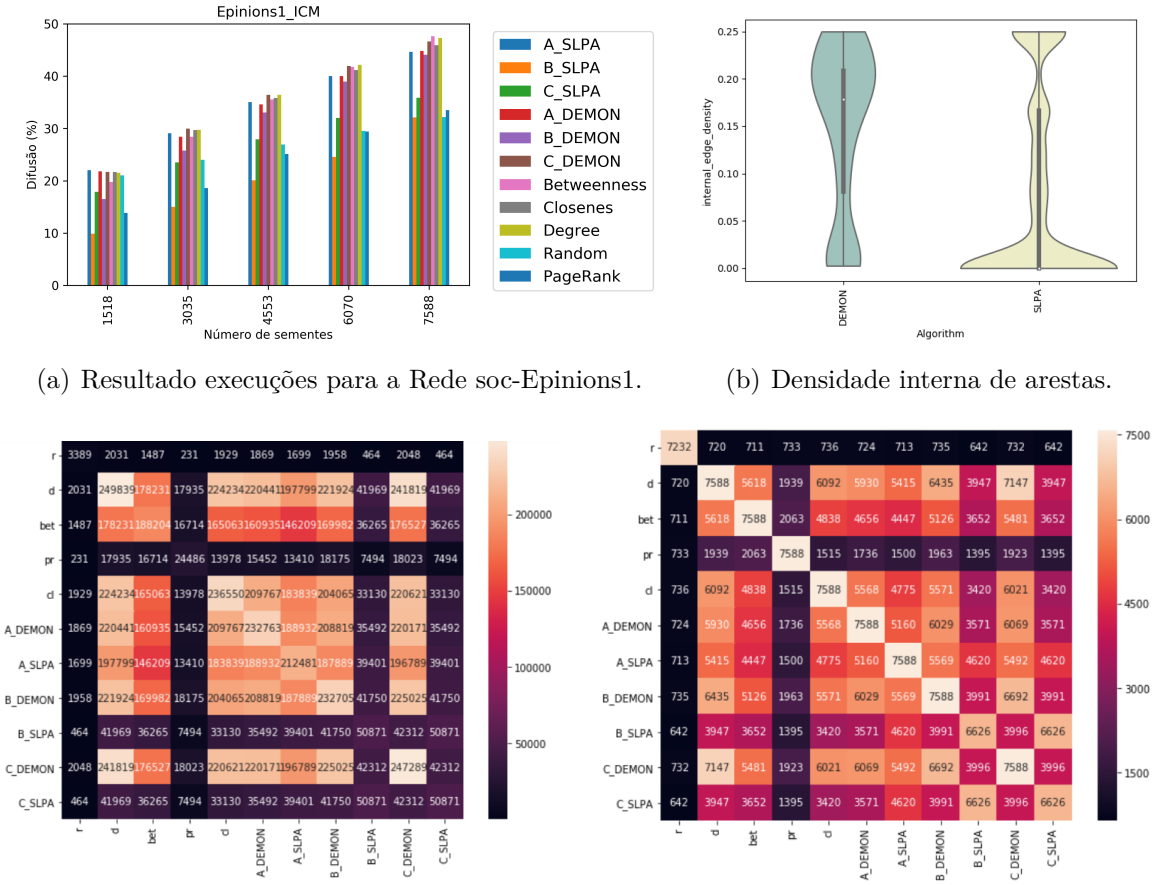
(b) Densidade interna de arestas.

(c) Relação de Arestas.

(d) Relação de Sementes.

Figura 25: Resultados alcançados para a rede Polbooks sob o modelo ICM.

algoritmo Demon se deva a quantidade de comunidades mais densas em sua maioria. E que provavelmente as sementes em um ambiente com número elevado de conectividade. Uma vez que o critério este critério também se baseia em centralidade de grau.



(a) Resultado execuções para a Rede soc-Epinions1.

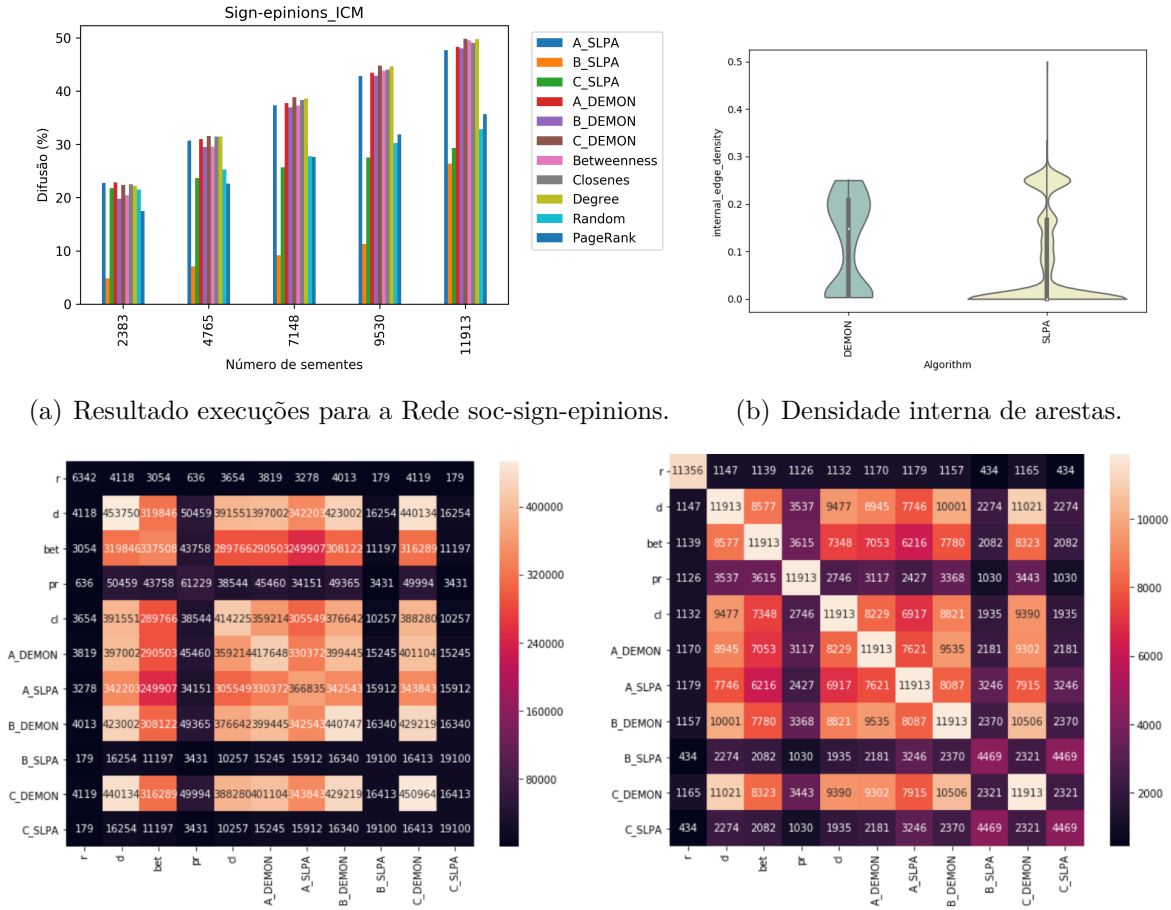
(b) Densidade interna de arestas.

(c) Relação de Arestas.

(d) Relação de Sementes.

Figura 26: Resultados alcançados para a rede soc-Epinions1 sob o modelo ICM.

Os resultados para a rede soc-Epinions1, como exibido na Figura 26, contabilizados o processo de difusão com 100% dos elementos contidos em cada subconjunto, alcança os melhores resultados com *Betweenness*, grau e C_Demon, respectivamente. O que faz com a topologia da rede propicie um eficiente processo de difusão utilizando elementos que são pontes. Analisando as relações entre as sementes e arestas, pode-se identificar que o subconjunto vencedor, possui também o menor subconjunto de arestas. Embora as diferenças entre a taxa de difusão entre eles no resultado geral normalizado seja de apenas um ponto percentual. Interessante observar que o comportamento dos critérios A_Demon e A_Slpa para todas as execuções são praticamente idênticos. Provavelmente porque ambos consideram apenas contribuições de acordo com o tamanho da comunidade e não em relação estrita a sobreposições. Quando considera-se o alcance da difusão para os critérios B e C sob o algoritmo SLPA provavelmente se deve ao número inferior de sementes em relação aos outros. A distribuição quanto a densidade interna de arestas nas comunidades é bem mais discrepante para o algoritmo SLPA, o que também pode ser uma justificativa pelo desempenho inferior.



(a) Resultado execuções para a Rede soc-sign-epinions.

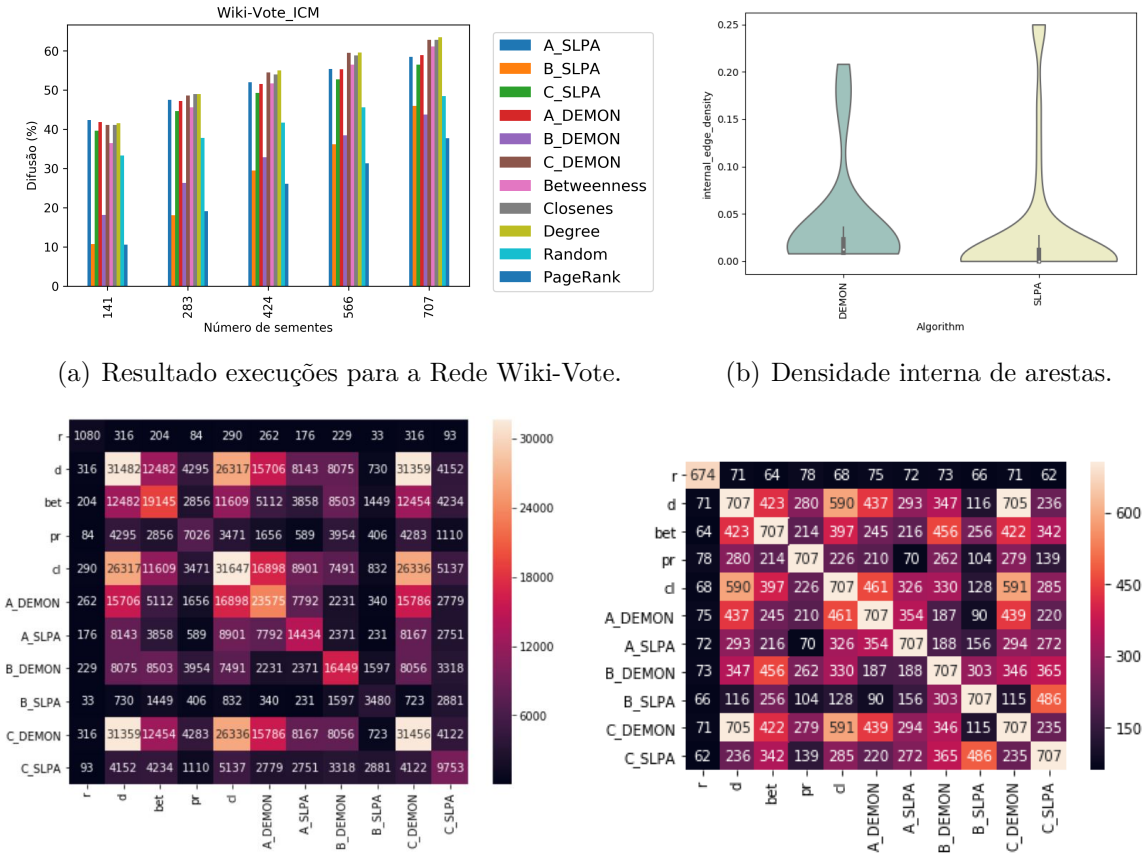
(b) Densidade interna de arestas.

(c) Relação de Arestas.

(d) Relação de Sementes.

Figura 27: Resultados alcançados para a rede soc-sign-epinions sob o modelo ICM.

O comportamento desta rede é muito similar ao da rede soc-Epinions1. Provavelmente porque ambas as redes estão inseridas no mesmo contexto, ambas são redes que mede a opinião de clientes sobre produtos em uma rede social. Na normalização dos dados a métrica de grau e o critério C_Demon empatam em primeiro lugar. Todavia, observando os dados brutos, o segundo tem uma margem insignificante de melhoria no processo. Muito provavelmente como pode ser observado pela Figura 27(b) a maioria das comunidades de tamanho relevante acima da média das outras, assim pode-se assumir que o alto grau das sementes nas interseções dessa região são responsáveis pelo sucesso deste critério. Ao contrário da distribuição interna da densidade de arestas entre as comunidades são discrepantes, bem como a linha fina, indica muitas pequenas comunidades com densidade de grau muito alto em relação a média. O que pode significar que sementes não alcancem estas comunidades, ou sejam tão densas que a cascata de difusão falhe.



(a) Resultado execuções para a Rede Wiki-Vote.

(b) Densidade interna de arestas.

(c) Relação de Arestas.

(d) Relação de Sementes.

Figura 28: Resultados alcançados para a rede Wiki-Vote sob o modelo ICM.

Na Figura 28(a) pode-se perceber uma similaridade entre o comportamento tanto das métricas quanto dos critérios com as duas redes acima. Apesar de outro contexto, a Wiki-Vote também é uma rede de confiança. O melhor desempenho neste caso, é da métrica de grau, seguida respectivamente, pela métrica CC e o critério C_Demon. Vale ressaltar que das redes do estudo é aquela com o grau mais alto. Assim, o resultado similar entre a métrica de centralidade de grau e o C_Demon não é surpresa. Bem como observar que as arestas e sementes em comum entre elas são basicamente as mesmas. Voltando no exemplo de *marketing*, a métrica CC é a que possui maior número de arestas, dessa forma seria mais barato utilizar o critério C_Demon para difundir uma marca ou produto.

6.1.2 Redes Sintéticas

A Tabela 6.1.2 mostra os resultados para as redes LFR com $\mu_u = 0.1$.

LFR_ICM 0.1

	A_SLPA	B_SLPA	C_SLPA	A_DEMON	B_DEMON	C_DEMON	BC	CC	DC	Random	PageRank
LFR1	0.80	0.88	0.86	0.86	0.92	0.98	0.98	0.95	0.99	0.86	1.00
LFR2	0.84	0.71	0.75	0.83	0.93	1.00	0.99	0.98	0.99	0.87	1.00
LFR3	0.78	0.65	0.59	0.79	0.95	1.00	0.95	0.98	0.98	0.82	0.97
LFR4	0.75	0.68	0.66	0.80	0.93	0.99	0.98	0.96	0.96	0.83	1.00
LFR5	0.84	0.76	0.74	0.87	0.94	1.00	0.98	0.97	0.98	0.88	0.98
LFR6	0.86	0.84	0.89	0.86	0.95	1.00	0.99	0.98	1.00	0.88	0.98
LFR7	0.77	0.82	0.86	0.86	0.92	0.97	0.98	0.96	1.00	0.82	0.99
LFR8	0.74	0.79	0.84	0.84	0.92	0.96	0.99	0.97	0.99	0.83	1.00
Total	6.38	6.13	6.19	6.71	7.46	7.90	7.84	7.75	7.89	6.79	7.92

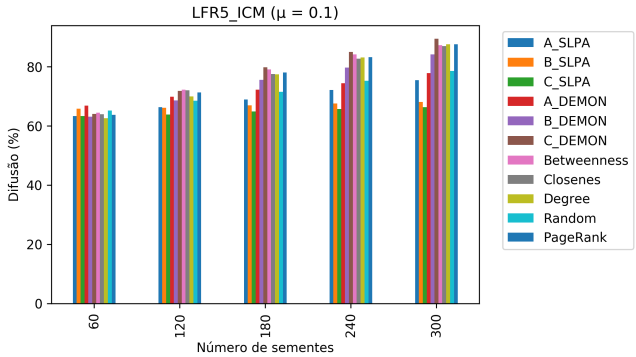
Tabela 6: Resultados para as redes sintéticas com execução sob o Modelo *Independent Cascade*, cujo valor do limite, (*threshold* = 0.1), o mesmo utilizado na biblioteca NDlib.

Como já observado anteriormente, o algoritmo de detecção de comunidades, SLPA, quando executado sobre as redes sintéticas com $\mu_u = 0,2$, identificaram a rede como uma única comunidade. A não existência de estrutura de comunidades, e consequentemente sobreposições, fizeram com que o experimento fosse descartado. Dos resultados retornados a partir da configuração do parâmetro $\mu_u = 0,1$, novamente o algoritmo Slpa não alcançou o número de sementes (10%) para os critérios B e C nas redes LFR2, LFR3, LFR4 e LFR5. Ao contrário, as sementes encontradas estão em torno de 1% a 40% dos 10% esperados. Assim, apenas as redes LFR1, LFR6, LFR7 e LFR8, seguindo esta definição, se tornaram elegíveis para avaliação. Entretanto, avaliando os intervalos entre os dados provenientes da Tabela 6.1.2, eles dizem que o comportamento para todos os subconjuntos de sementes, independente de seu tamanho são similares. Excetuando-se um caso quando B_Slpa, na rede LFR2, não alcança nem 2% de difusão.

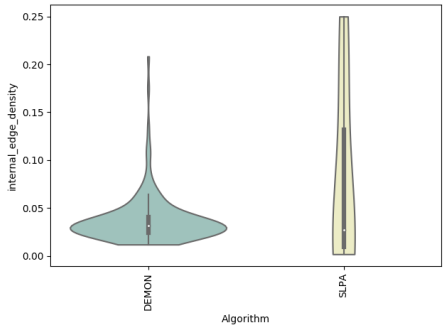
Para os critérios B e C do SLPA e sementes aleatórias os subconjuntos apresentam baixa interseção com os outros subconjuntos. Diante do exposto, apenas os resultados para as redes LFR5 e LFR6 serão apresentados. A LFR5, cujo quantitativo de sementes obtidas pela SLPA é de apenas 33, contra contra 300 dos outros subconjuntos, justificando seu baixo desempenho. Já a rede LFR6 que contempla a o número de sementes para todos os métodos utilizados.

Dessa forma, é possivel visualizar que a discrepância entre os subconjuntos não retira a eficiência dos critérios A e B do SLPA. Possivelmente as sementes selecionadas pelos critérios B e C do SLPA obtiveram um bom resultado no processo de difusão devido a similaridade das comunidades encontradas, ou seja, entre tamanho e distribuição de densidade interna de arestas. Diferentemente, o critério C_Demon, alcança o melhor resultado para esta rede, embora seu alcance possa ser prejudicado por comunidades pequenas, cuja densidade interna de arestas pode não contemplar sementes e/ou não ser alcançadas por elas.

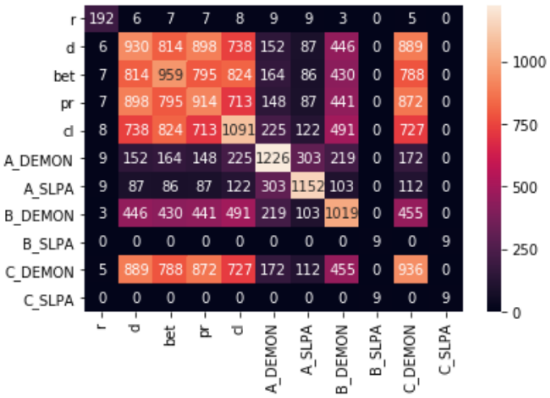
Quanto a rede LFR6, no geral, seu comportamento é semelhante as outras redes. O melhor resultado para a mesma identificado na Tabela 6.1.2 vai para C_Demon e grau de centralidade. De maneira que é o resultado é coerente, pois o critério em questão tem agregada ao método de seleção, o grau para as sementes que estão em sobreposição. Da mesma forma que a rede LFR5, a eficiência do critério C_Demon pode ser prejudicado por pequenas comunidades densas.



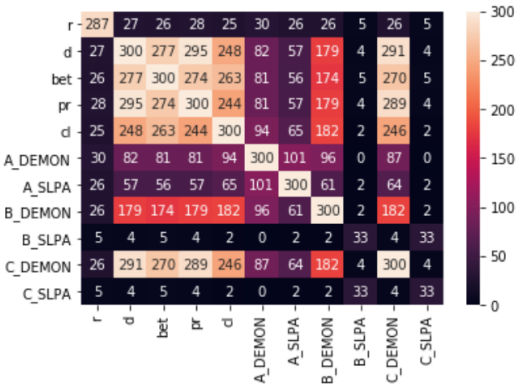
(a) Resultado execuções para a Rede LFR5.



(b) Densidade interna de arestas.

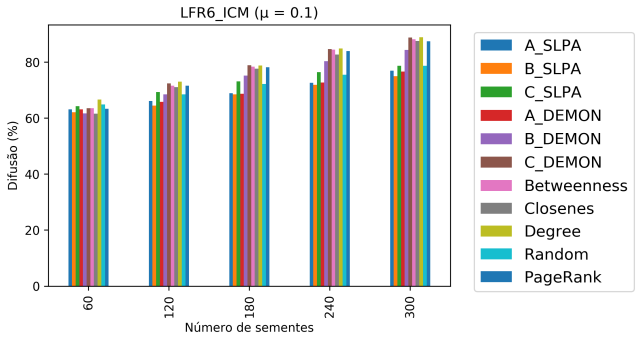


(c) Relação de Arestas.

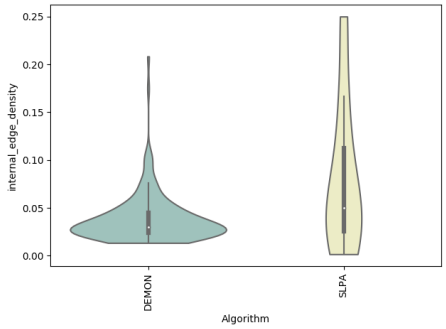


(d) Relação de Sementes.

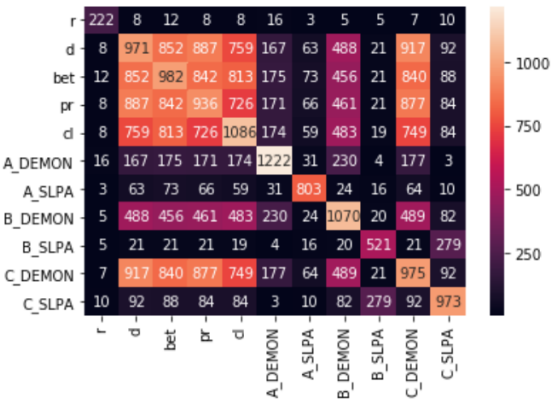
Figura 29: Resultados alcançados para a rede sintética LFR5 sob o modelo ICM.



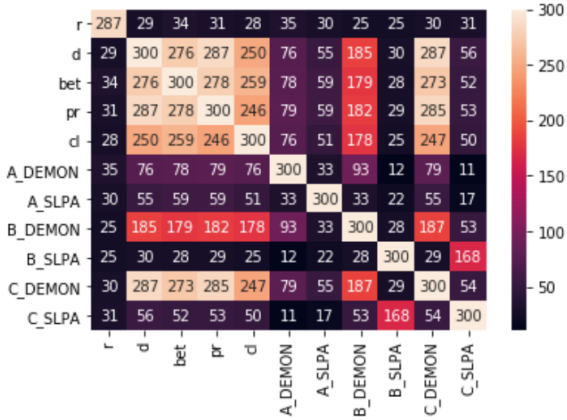
(a) Resultado execuções para a Rede LFR6.



(b) Densidade interna de arestas.



(c) Relação de Arestas.



(d) Relação de Sementes.

Figura 30: Resultados alcançados para a rede sintética LFR6 sob o modelo ICM.

Vale ressaltar que, os melhores resultados para os critérios que utilizam de sobreposição, estão nas redes do tipo *benchmark*. Onde as configurações possuem menor expoente de lei de potência tanto para a rede quanto para as comunidades.

6.2 Resultados da Difusão sob o Modelo LTM

6.2.1 Redes Reais

Esta subseção apresenta os resultados para o processo de difusão considerando o modelo LTM, cujo parâmetro de limite para as arestas é, $threshold = 0,3$.

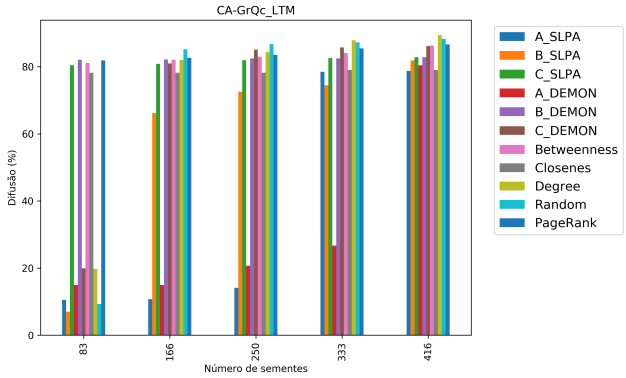
Para melhor relacionar os resultados das execuções sob o modelo *Threshold*, as imagens já apresentadas nos resultados para o modelo *Independent Cascade* serão também exibidos nesta Seção. Ou seja, os gráficos de densidade interna de arestas (resultado da detecção de comunidades) e as imagens da correlação entre os subconjuntos de arestas e sementes resultante dos subconjuntos de sementes selecionadas para as redes.

A Tabela 6.2.1, mostra os resultados gerais para métricas e critérios das execuções com o total de sementes dos subconjuntos. Igualmente ao modelo ICM, as redes cujos resultados serão apresentados são: CA-GrQc, CA-HepTH, moreno_innovation, Polbooks, soc-Epinions1, soc-sign- epinions e Wiki-Vote.

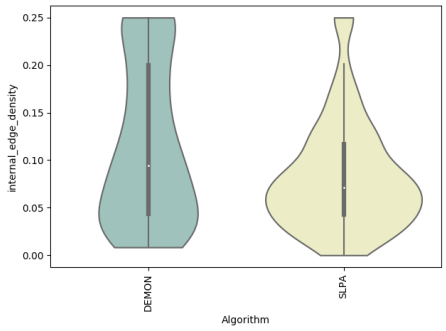
	A_SLPA	B_SLPA	C_SLPA	A_DEMON	B_DEMON	C_DEMON	BC	CC	DC	Random	PageRank
CA-GrQc	0.88	0.92	0.93	0.90	0.93	0.96	0.97	0.88	1.00	0.99	0.97
CA-HepTh	0.96	0.98	0.97	0.96	0.96	0.97	0.98	0.96	0.98	1.00	1.00
Dolphins	1.00	0.00	0.00	1.00	0.48	1.00	1.00	1.00	1.00	1.00	1.00
Football	0.14	0.00	0.00	0.14	0.00	0.00	1.00	1.00	1.00	0.20	1.00
Jazz	1.00	0.00	0.00	1.00	0.34	1.00	1.00	1.00	1.00	0.13	1.00
karate	1.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Moreno Innovation	1.00	0.37	0.96	1.00	0.91	0.93	0.88	0.97	0.93	0.96	0.12
Polbooks	1.00	0.11	0.11	1.00	0.48	1.00	1.00	1.00	1.00	1.00	1.00
Epinions1	1.00	0.70	0.70	1.00	1.00	1.00	1.00	1.00	1.00	0.24	0.53
Sign-epinions	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.23	0.62
Wiki-Vote	1.00	0.36	0.99	0.99	0.65	1.00	0.97	1.00	1.00	0.18	0.45
Total	9.98	4.44	5.66	9.99	7.75	9.86	10.80	10.81	10.91	6.93	8.69

Tabela 7: Resultados para as redes reais com execução sob o Modelo *Threshold*, cujo valor do limite (*threshold=0.3* próximo aos 0.25 que a biblioteca NDlib utiliza.

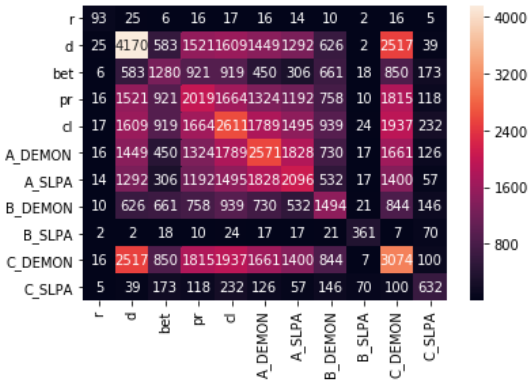
As execuções para a rede CA-GrQc sob o modelo Threshold são exibidos na Figura 31. O comportamento da difusão, pode ser verificado na Figura 31(a), é difícil identificar o melhor subconjunto de sementes. Na última execução, tanto critérios como métricas alcançam resultados similares. Entretanto da primeira execução para a últimas, os resultados são muito diferentes. O melhor resultado alcançado no processo de difusão é o grau, na última execução. Por outro lado, na primeira execução com 20% do total deste subconjunto de sementes seu alcance é muito baixo. Interessante é que o critério que utiliza desta métrica, além da sobreposição, possui resultado muito semelhante, em ambos os extremos das execuções, primeira e última execução. O que se pode inferir, neste caso, que as sementes com maior grau, mas nas sobreposições foram mais eficientes no processo de difusão. Por outro lado, o resultado da última execução, para a medida de centralidade de grau, se deva mais ao modelo de difusão, bem como ao limite das arestas da rede, do que das estratégias utilizadas. Retornando a exemplos anteriores, tal como, restrições de custo para um processo de difusão, o subconjunto de arestas do critéro C_Demon na Figura 31(c) teria custo mais baixo que o subconjunto que utiliza medida de grau.



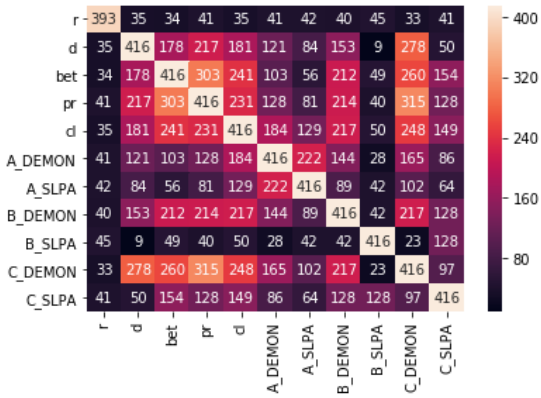
(a) Resultado execuções para a Rede CA-GrQc.



(b) Densidade interna de arestas.



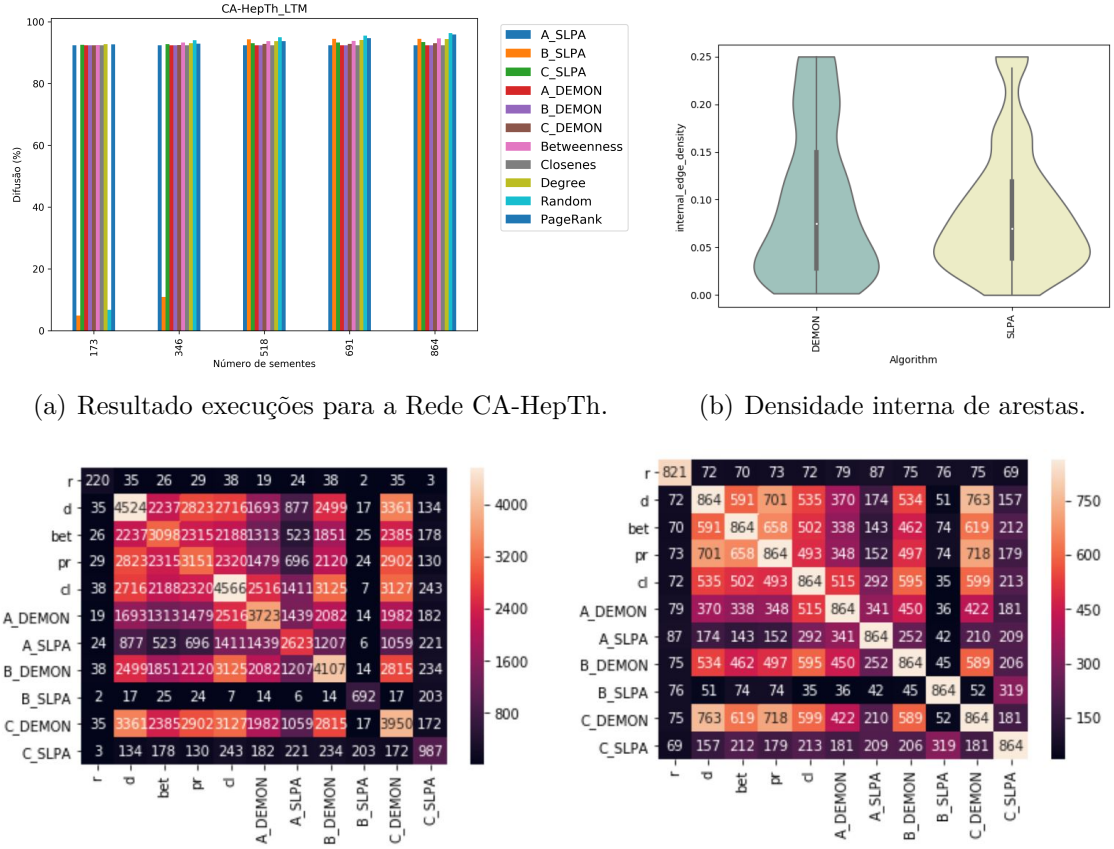
(c) Relação de Arestas.



(d) Relação de Sementes.

Figura 31: Resultados alcançados para a rede GA-GrQc sob o modelo LTM.

A Figura 32 mostra os resultados para a rede HepTh sob o modelo *Threshold*. Para esta rede, como mostra a Figura 32(a), o total do conjunto de sementes conta com 864 sementes. A primeira execução com 10% (173) sementes, os resultados são praticamente



(a) Resultado execuções para a Rede CA-HepTh.

(b) Densidade interna de arestas.

(c) Relação de Arestas.

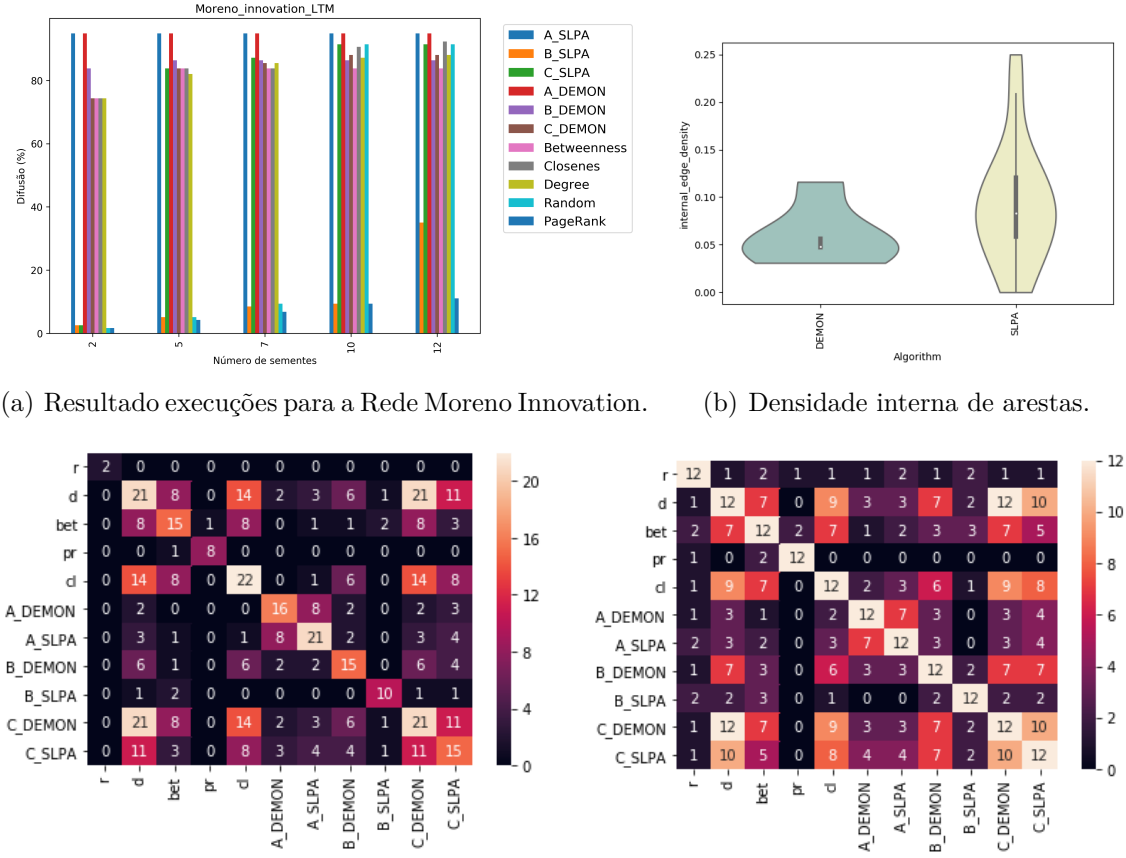
(d) Relação de Sementes.

Figura 32: Resultados das execuções para a rede HepTh sob o modelo LTM.

idênticos para critérios e medidas de centralidade, a exceção diz respeito aos subconjunto B_Slpa e sementes aleatórias. Os resultados para estes subconjuntos, neste caso, são extremamente inferior aos outros. Vale observar que a similaridade entre os subconjuntos de arestas para o subconjunto do critério B_Slpa é dispare. Dessa forma, uma questão surge neste contexto. Quais são os elementos que acrescidos na terceira execução faz com que o critério tenha resultado superior aos outros subconjuntos e similar às sementes aleatórias?

Na Figura 33(a), para a rede moreno_innovation, um outro comportamento é notado. Para todas as execuções os critérios A_Slpa e A_Demon se sobressaem, mantendo o alcance superior aos outros métodos desde a execução com 20% do total do subconjunto de sementes, mesmo que ambos tenham apenas 8 arestas em comum. Muito provavelmente o alcance da difusão para estes critérios deve estar ligado a dinâmica do modelo de difusão, bem como ao limite de configuração do modelo (0,3), pois a similaridade entre os subconjuntos de arestas e sementes em sua maioria são muito diferentes. Dessa forma, pode concluir que cada critério possui sementes com características específicas importantes para um processo de difusão.

Para a rede Polbooks, como é mostrado na Figura 34(a), utilizando apenas uma se-



(a) Resultado execuções para a Rede Moreno Innovation.

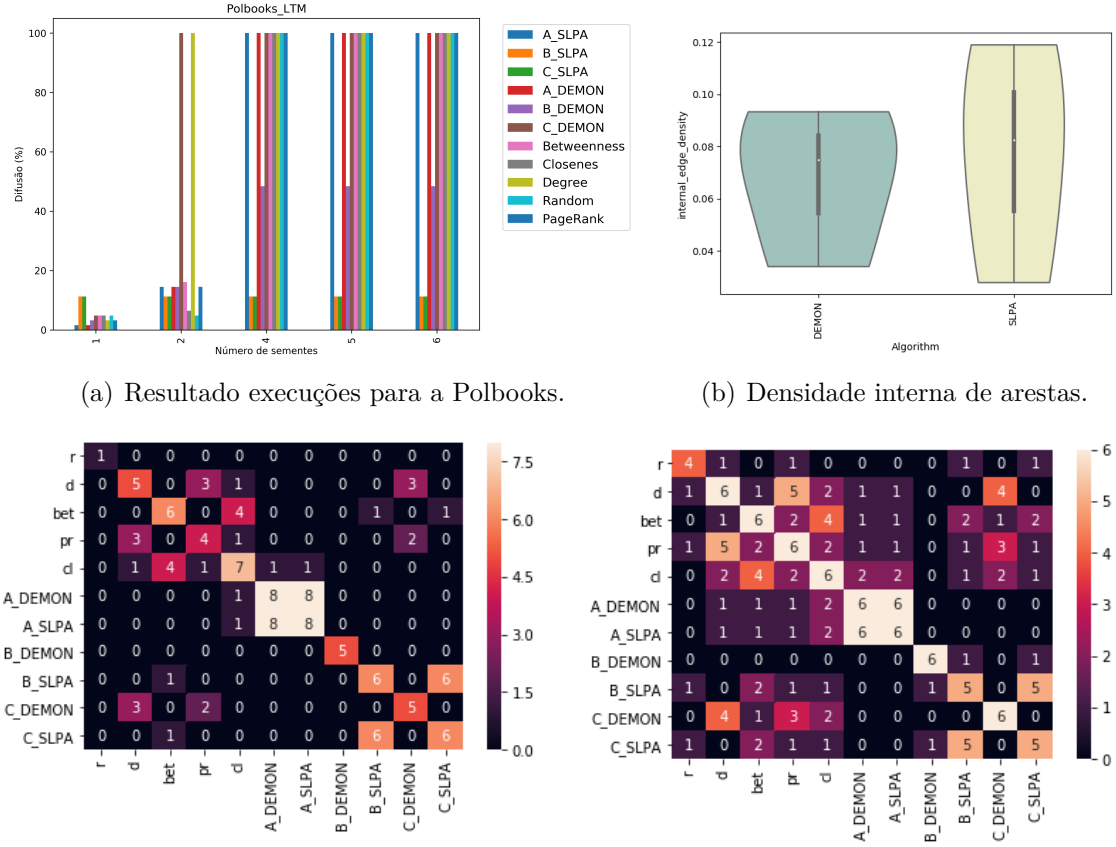
(b) Densidade interna de arestas.

(c) Relação de Arestas.

(d) Relação de Sementes.

Figura 33: Resultados das execuções para a rede Moreno Innovation sob o modelo LTM.

mentes, os critérios B e C do SLPA apresentam relevância em relação aos outros. Embora com o acréscimo de mais uma semente ao subconjunto, o resultado seja muito inferior a C_Demon e centralidade de grau que já alcançam nesta execução 100% de *nós* influenciados na rede. De todos os outros métodos que alcançam 100% já na terceira execução (4 sementes) de difusão na rede, apenas um outro critério, B_Demon tem resultado também inexpressivo e os critérios B e C do SLPA que mantêm o mesmo resultado com 1 e 6 sementes. Observando o subconjunto de arestas para o critério B e C do SLPA é verificado que ambos possuem apenas uma semente compartilhada com o subconjunto da medida de centralidade *betweenness*.



(a) Resultado execuções para a Polbooks.

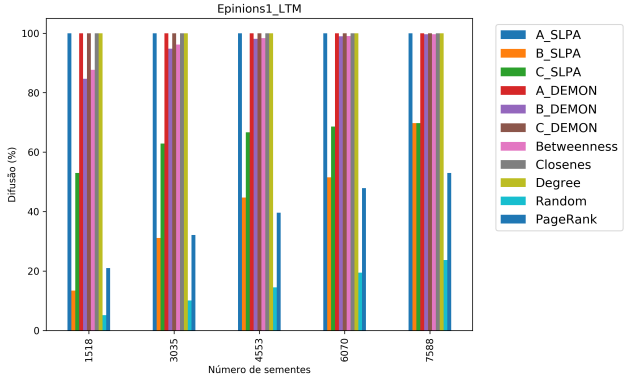
(b) Densidade interna de arestas.

(c) Relação de Arestas.

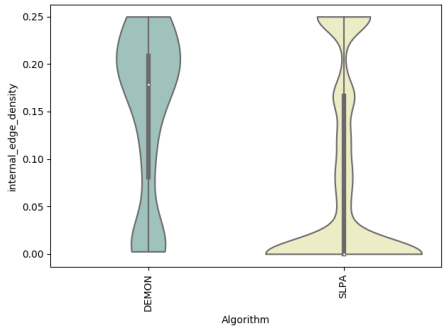
(d) Relação de Sementes.

Figura 34: Resultados das execuções para a rede Polbooks para o modelo LTM.

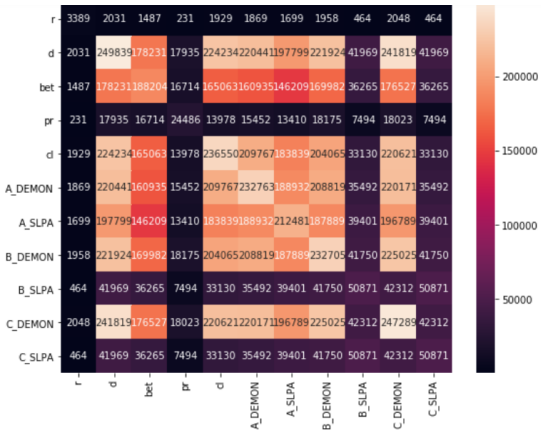
Igualmente aos resultados apresentados, soc-Epinions1 e soc-sign-epinions, nas Figuras 35(a) e 36(a) sob o modelo *Linear Threshold*, o comportamento para estas redes se mostra similar ao modelo *Independent Cascade*. Vale observar que o critério B_Slpa possui o pior desempenho entre os outros critérios na primeira execução, mas se equipara ao ao C_Slpa na última delas. Seu comportamento é linear para a rede Epinions1, embora para a sign-epinions no resultado inferior a 20% de difusão na primeira execução, ele alcança resultado inferior apenas para o critério A_SLPA. Independente dos modelos utilizados, possivelmente os algoritmos de detecção de comunidades estão influenciando os resultados apresentados.



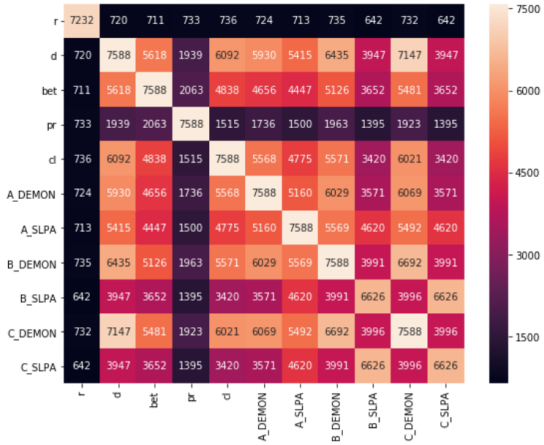
(a) Resultado execuções para a soc-Epinions1.



(b) Densidade interna de arestas.

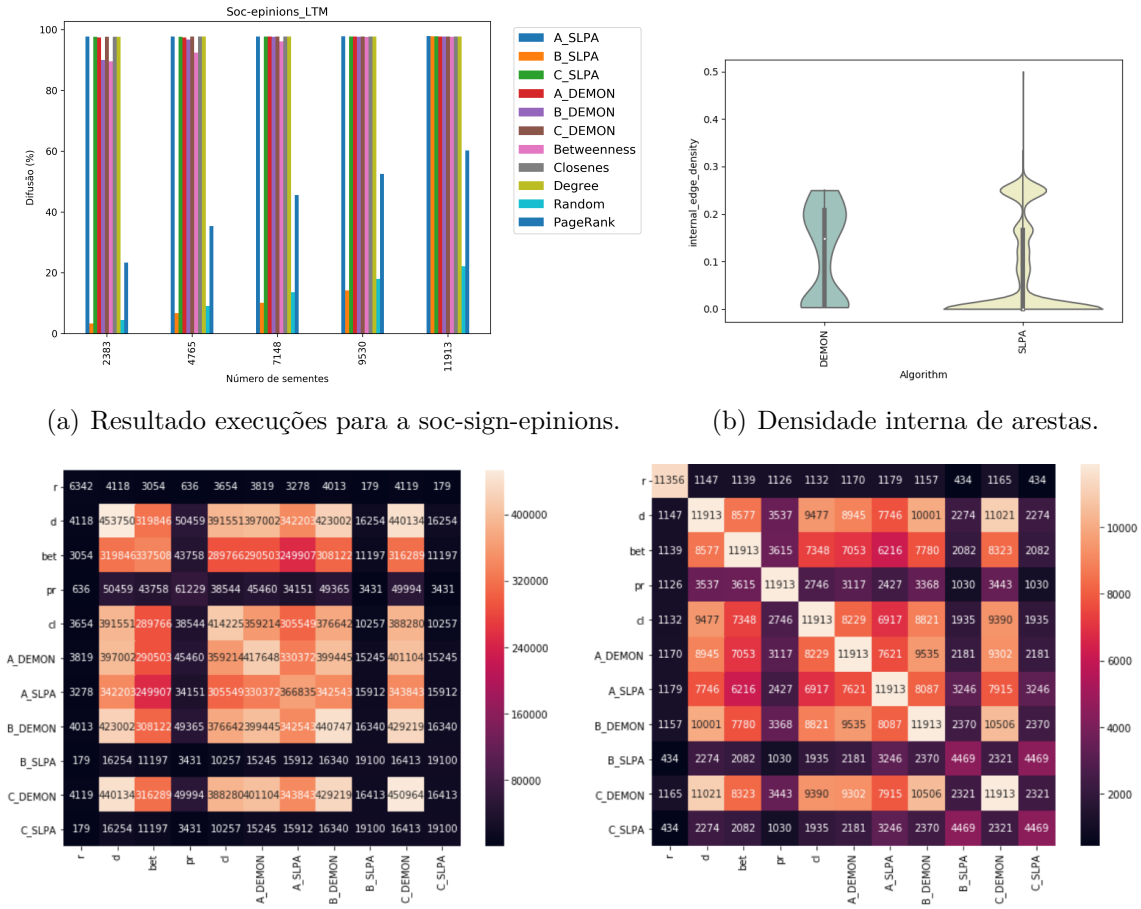


(c) Relação de Arestas.



(d) Relação de Sementes.

Figura 35: Resultados das execuções para a rede soc-Eoinions1 sob o modelo LTM.



(a) Resultado execuções para a soc-sign-epinions.

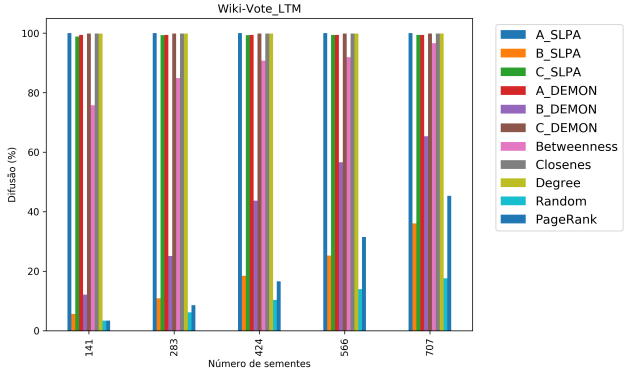
(b) Densidade interna de arestas.

(c) Relação de Arestas.

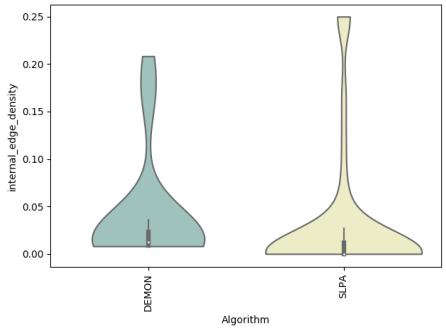
(d) Relação de Sementes.

Figura 36: Resultados das execuções para a rede soc-sign-epinions sob o modelo LTM.

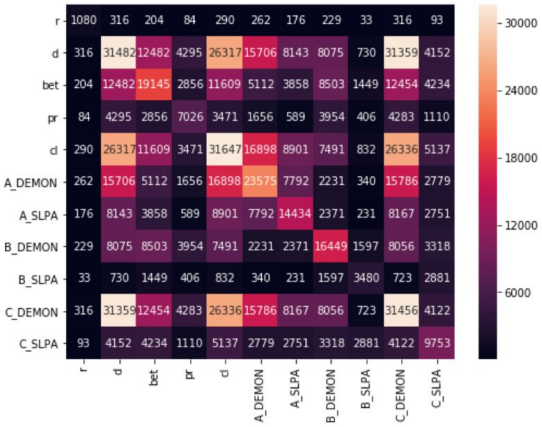
Como exibido na Figura 37(a), o desempenho dos critérios B, tanto o derivado do algoritmo Demon quanto do algoritmo SLPA possuem desempenho muito inferior aos outros critérios e superior apenas a métrica de centralidade *PageRank* e o subconjunto de sementes aleatórias. Para os demais métodos, apenas o *Betweennes* não alcança a taxa de difusão similar aos outros. Vale dizer que, nos dados observados para todas as execuções (intervalo de 10% a 100% do conjunto de sementes), o critério A_Slpa obteve alcance de 100% em todos. Na Figura 37(b) é observada a distribuição de densidade interna de arestas geradas na detecção de comunidades. É observado que a rede possui um número expressivo de pequenas comunidades densas. Dessa forma não é possível identificar se as sementes geradas para os critérios B estão ou não na interseção destas comunidades. Se não, recai-se novamente na alegação de (EASLEY, 2010) de que grupos densos quebram uma cascata.



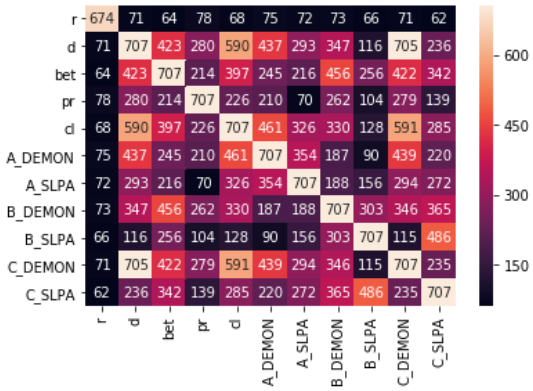
(a) Resultado execuções para a wiki-vote.



(b) Densidade interna de arestas.



(c) Relação de Arestas.



(d) Relação de Sementes.

Figura 37: Resultados das execuções para a rede Wiki-Vote para o modelo LTM.

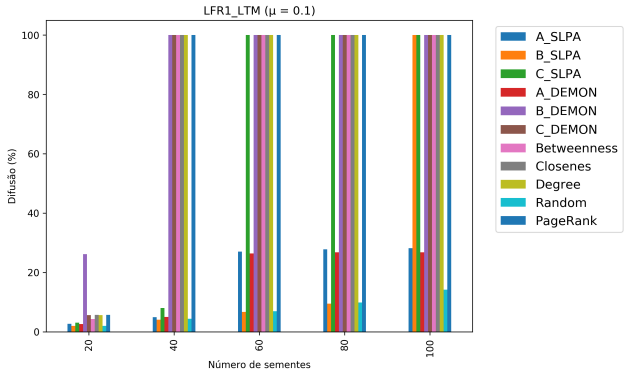
6.2.2 Redes Sintéticas

	A_SLPA	B_SLPA	C_SLPA	A_DEMON	B_DEMON	C_DEMON	Bc	CC	DC	Random	PageRank
LFR1	0.28	1.00	1.00	0.27	1.0	1.0	1.0	1.0	1.0	0.14	1.0
LFR2	0.26	0.01	0.01	0.26	1.0	1.0	1.0	1.0	1.0	1.00	1.0
LFR3	0.27	0.02	0.02	0.27	1.0	1.0	1.0	1.0	1.0	1.00	1.0
LFR4	1.00	0.05	0.05	1.00	1.0	1.0	1.0	1.0	1.0	1.00	1.0
LFR5	0.28	0.01	0.01	1.00	1.0	1.0	1.0	1.0	1.0	0.12	1.0
LFR6	1.00	1.00	1.00	0.53	1.0	1.0	1.0	1.0	1.0	0.13	1.0
LFR7	0.26	1.00	1.00	1.00	1.0	1.0	1.0	1.0	1.0	0.13	1.0
LFR8	0.26	1.00	1.00	1.00	1.0	1.0	1.0	1.0	1.0	0.17	1.0
Total	3.61	4.09	4.09	5.33	8.0	8.0	8.0	8.0	8.0	3.69	8.0

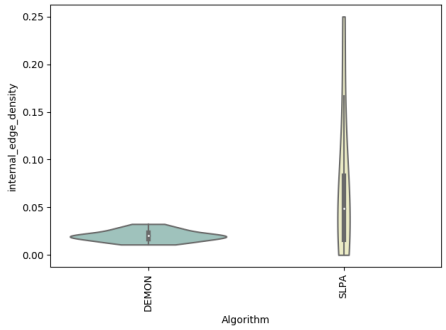
Tabela 8: Resultados para as redes sintéticas com execução sob o Modelo *Threshold*, cujo valor do limite (*threshold=0.3*) próximo aos 0.25 que a biblioteca NDlib utiliza.

Os resultados apresentados para o processo de difusão considera o modelo *Linear Threshold* quando o limite estabelecido para cada aresta decidir se é influenciado ou não pelo comportamento de seus vizinhos é de $0,3$.

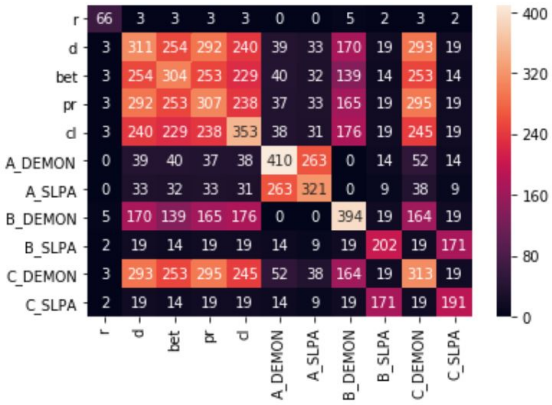
Nas Figuras 38, 39, 40 e 41 são exibidos os resultados para execuções do experimento para as redes LFR1, LFR6, LFR7 e LFR8. As mesmas foram selecionadas para exibição dos reultados porque alcançaram o número (10% do total da rede) de sementes, para todos subconjuntos após serem aplicados os algoritmos de detecção de comunidades e os algoritmos para seleção baseada nos critérios.



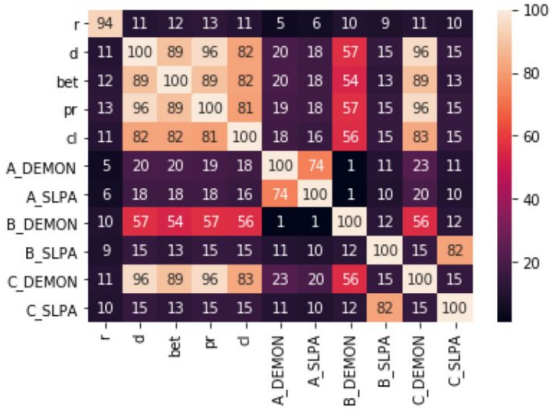
(a) Resultado execuções para a Rede LFR1.



(b) Densidade interna de arestas.



(c) Relação de Arestas.

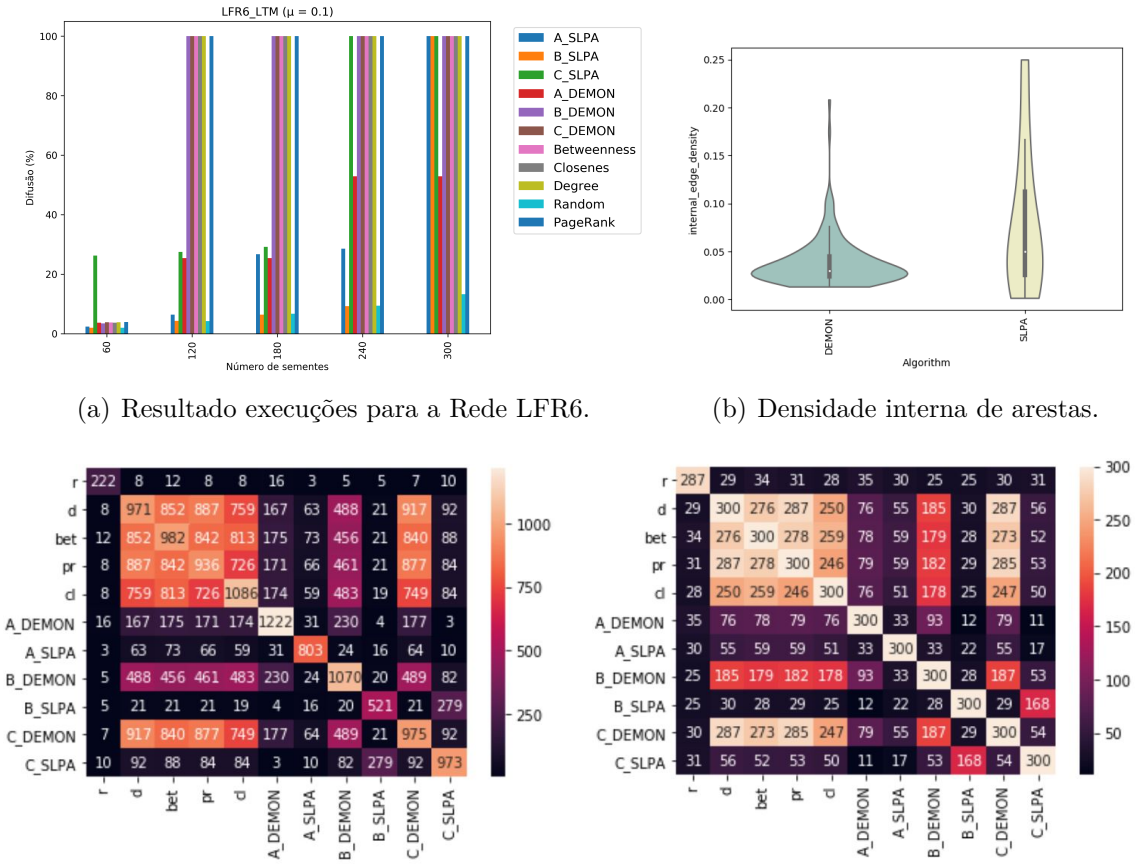


(d) Relação de Sementes.

Figura 38: Resultados alcançados da difusão sob o modelo *Threshold* quando $\mu_u = 0.1$ e o limite é 0.3 para a rede LFR1.

A similaridade observada no comportamento do experimento para todas as redes na execução com 20% das sementes é praticamente inexpressivo. A partir das execuções com 40% dos subconjuntos o alcance é total na maioria das situações. Quando as configurações da rede são $\tau_1 = 2,0$ e $\tau_1 = 1,2$ (ver Tabela 2 na seção 5.1), a execução com 10% das sementes para os critérios B_Demon e C_Slpa, nas redes LFR1 e LFR6 respectivamente, possuem comportamento expressivo. Neste caso, a restrição de se utilizar poucas sementes para um processo de difusão, torna estes critérios eficientes.

Na rede LFR7, Figura 40(a), na execução com 20% dos subconjuntos de sementes



(a) Resultado execuções para a Rede LFR6.

(b) Densidade interna de arestas.

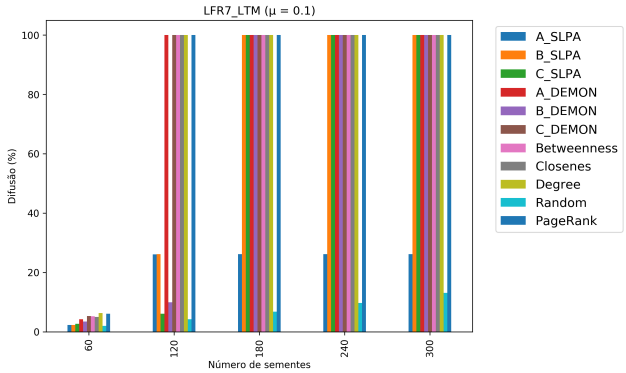
(c) Relação de Arestas.

(d) Relação de Sementes.

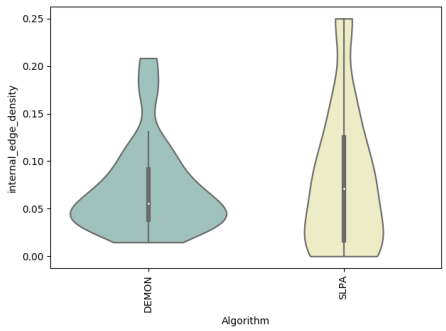
Figura 39: Resultados alcançados da difusão sob o modelo *Threshold* quando $\mu_u = 0.1$ e o limite é 0.3 para a rede LFR6.

os resultados para todos os métodos são ineficientes. Embora, com a execução com 40%, os critérios A_Demon e C_Demon tem resultados equivalestes as medidas de centralidade. Mesmo considerando que o número de arestas e sementes destes dois subconjuntos tenham menos elementos em comum do que o as medidas de centralidade.

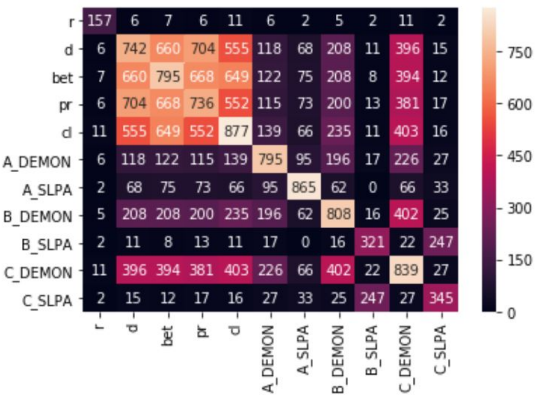
Na rede LFR8, Figura 41(a), este comportamento é observado para os critérios C_Slpa e C_Demom, com número também inferior de similaridade entre os subconjuntos de arestas e sementes.Provavelmente este comportamento, bem como para a rede LFR7, está relacionado ao modelo de difusão.



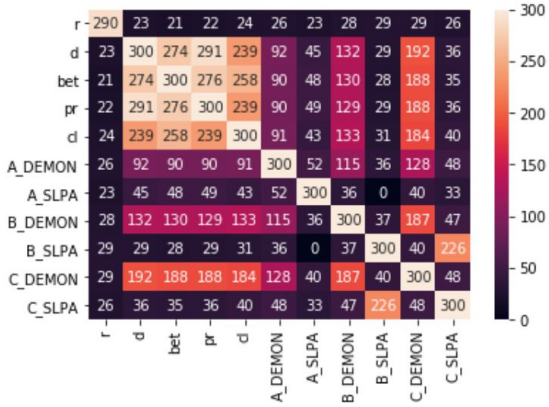
(a) Resultado execuções para a Rede LFR7.



(b) Densidade interna de arestas.

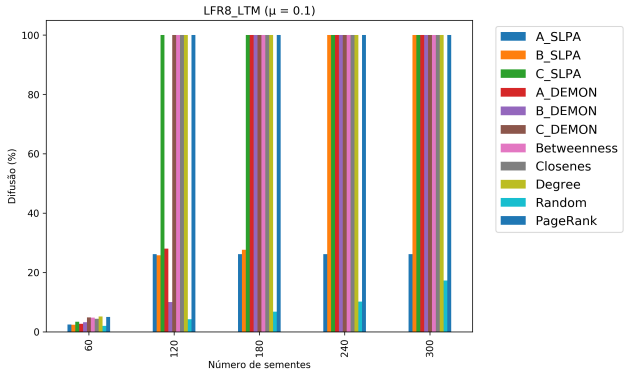


(c) Relação de Arestas.

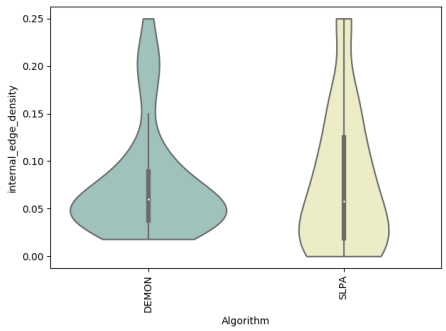


(d) Relação de Sementes.

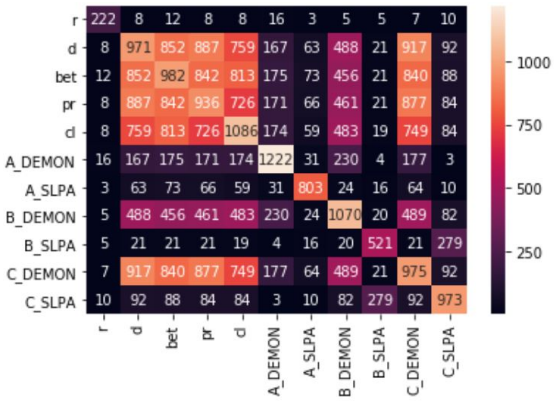
Figura 40: Resultados alcançados da difusão sob o modelo *Threshold* quando $\mu_u = 0.1$ e o limite é 0.3 para a rede LFR7.



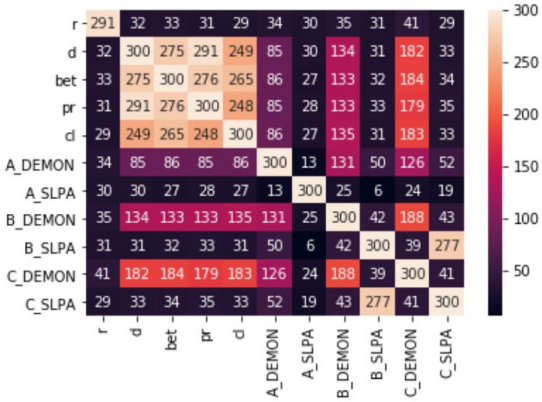
(a) Resultado execuções para a Rede LFR8.



(b) Densidade interna de arestas.



(c) Relação de Arestas.



(d) Relação de Sementes.

Figura 41: Resultados alcançados da difusão sob o modelo *Threshold* quando $\mu_u = 0.1$ e o limite é 0.3 para a rede LFR8.

7 Conclusão e Trabalhos Futuros

O estudo sobre influência social tem atraído a atenção de diversos domínios há décadas, tal como sociologia e marketing, mas tem se tornado também popular entre cientistas da computação. Neste contexto, o estudo tem sido facilitado pela massiva quantidade de dados disponíveis como resultado do avanço tecnológico e as emergentes redes sociais. O entendimento e compreensão dos fenômenos percebidos, sobretudo a difusão de informação em redes sociais online, o interesse de profissionais ligados ao marketing devido a popularização do mercado online, entre outros. Este cenário também fornece possíveis maneiras de identificar percursos por onde um comportamento pode ser difundido. E mais, quais seriam os elementos na estrutura de uma rede social mais eficientes.

Neste sentido, este trabalho buscou identificar elementos na estrutura de comunidades de uma rede, mais especificamente em partições com comunidades sobrepostas, visto que são comuns em redes reais. A ideia vem da hipótese de que elementos em sobreposição de comunidades são mais eficientes em propagar um comportamento, haja vista que podem influenciar seus pares em mais de um grupo.

Tendo em vista os resultados encontrados, elementos em sobreposições de grupos não garantem, em muitos casos, a eficiência no processo de difusão. Após exaustivos experimentos, percebe-se que muitas variáveis podem ser exploradas, tal como um estudo mais acurado das comunidades, parâmetros dos métodos de detecção e mesmo explorar outros métodos. O que não refuta o fato de que nos experimentos realizados, sob as condições descritas, algumas métricas de centralidade obtiveram resultados mais eficientes.

Outro fato observado é que para redes reais, elencadas para discussão, o algoritmo Demon detectou um número menor de comunidades que o SLPA, e o contrário quando se tratou das execuções nas redes sintéticas.

No trabalho de (VIEIRA, 2019), um problema apontado pelos autores é de que o processo de algoritmos que detectam comunidades com sobreposição pode interferir mais nos resultados do que a estrutura da rede. No trabalho de (ZHOU, 2017), embora se trate de agrupamento de arestas, o mesmo problema é colocado, ou seja, a dificuldade em detectar grupos reais. Os autores apontam ainda a possibilidade de que seja encontrado um número excessivo de sobreposições.

Restrições impostas pelo uso das bibliotecas mencionadas, tiveram fatores que limitaram os experimentos, tal como a utilização de algoritmos de detecção de comunidades cujo parâmetro obrigatória é apenas a rede. Poucos algoritmos oferecem esta característica. Embora, o algoritmo LFM o ofereça, retornou erro em algumas redes em particular, sendo assim descartado.

O trabalho contribui com o referencial teórico que cobre os tópicos abordados. No que tange a contribuições metodológicas, o presente trabalho propicia a descoberta de novas variáveis passíveis de serem exploradas.

Sendo assim, trabalhos futuros podem ser definidos explorados, tais como: a utilização de outros algoritmos disponíveis na biblioteca, mesmo os que possuam uma gama de parâmetros que necessitem de ajustes; agregar a heurística de (CHEN, 2009) *Degree Discount* aos elementos encontrados nas sobreposições; realizar uma análise acurada das comunidades encontradas de maneira a que melhorias sejam propostas aos critérios definidos no trabalho; estudar o *framework* definido no trabalho por etapas, avaliando a sensibilidade de cada etapa de maneira a buscar possíveis alternativas de melhoria; propor um modelo híbrido para critério de seleção de sementes com sobreposição utilizando a centralidade de *betweenness*.

Referências

RASHOTTE, L. Social influence. *The Blackwell encyclopedia of sociology*, Wiley Online Library, 2007. Citado 2 vezes nas páginas 13 e 38.

KEMPE, D.; KLEINBERG, J.; TARDOS, É. Influential nodes in a diffusion model for social networks. In: SPRINGER. *International Colloquium on Automata, Languages, and Programming*. [S.l.], 2005. p. 1127–1138. Citado 2 vezes nas páginas 13 e 40.

GRANOVETTER, M. Threshold models of collective behavior. *American journal of sociology*, University of Chicago Press, v. 83, n. 6, p. 1420–1443, 1978. Citado 5 vezes nas páginas 13, 14, 38, 43 e 44.

NEWMAN, M. E. Spread of epidemic disease on networks. *Physical review E*, APS, v. 66, n. 1, p. 016128, 2002. Citado 2 vezes nas páginas 13 e 45.

ZHANG, H.; MISHRA, S.; THAI, M. T.; WU, J.; WANG, Y. Recent advances in information diffusion and influence maximization in complex social networks. *Opportunistic Mobile Social Networks*, USA: CRC Press, v. 37, n. 1.1, 2014. Citado 4 vezes nas páginas 13, 43, 44 e 45.

ROGERS, E. m. *Diffusionof innovations*. [S.l.]: Simon and Schuster, 2010. Citado 4 vezes nas páginas 13, 38, 42 e 43.

JOE, T. *Gaining momentum: Managing the diffusion of innovations*. [S.l.]: World Scientific, 2010. Citado na página 13.

GULATI, R. Alliances and networks. *Strategic management journal*, Wiley Online Library, v. 19, n. 4, p. 293–317, 1998. Citado 2 vezes nas páginas 13 e 15.

GREGORY, S. Finding overlapping communities in networks by label propagation. *New journal of Physics*, IOP Publishing, v. 12, n. 10, p. 103018, 2010. Citado 4 vezes nas páginas 13, 32, 33 e 35.

LIN, S.; HU, Q.; WANG, G.; PHILIP, S. Y. Understanding community effects on information diffusion. In: SPRINGER. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2015. p. 82–95. Citado 2 vezes nas páginas 13 e 59.

LEICHT, E.; NEWMAN, M. Community structure in directed networks. 2008. Citado 2 vezes nas páginas 14 e 31.

BAKSHY, E.; ROSENN, I.; MARLOW, C.; ADAMIC, L. The role of social networks in information diffusion. In: ACM. *Proceedings of the 21st international conference on World Wide Web*. [S.l.], 2012. p. 519–528. Citado na página 14.

LANCICHINETTI, A.; FORTUNATO, S.; KERTÉSZ, J. Detecting the overlapping and hierarchical community structure in complex networks. *New journal of physics*, IOP Publishing, v. 11, n. 3, p. 033015, 2009. Citado 3 vezes nas páginas 14, 33 e 56.

JALALI, M. S.; ASHOURI, A.; HERRERA-RESTREPO, O.; ZHANG, H. Information diffusion through social networks: The case of an online petition. *Expert Systems with Applications*, v. 44, p. 187 – 197, 2016. ISSN 0957-4174. Citado na página 14.

XIE, J.; KELLEY, S.; SZYMANSKI, B. K. Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur)*, ACM, v. 45, n. 4, p. 43, 2013. Citado na página 14.

EASLEY, D.; KLEINBERG, J. *Networks, crowds, and markets*. [S.l.]: Cambridge university press Cambridge, 2010. Citado 10 vezes nas páginas 14, 16, 19, 23, 38, 39, 43, 44, 45 e 78.

BARABASI, A. *Network Science*. Cambridge University Press, 2016. ISBN 9781107076266. Disponível em: <http://networksciencebook.com>. Citado 12 vezes nas páginas 16, 17, 18, 19, 20, 21, 22, 28, 30, 31, 32 e 45.

COHEN, R.; HAVLIN, S. *Complex networks: structure, robustness and function*. [S.l.]: Cambridge university press, 2010. Citado 7 vezes nas páginas 16, 17, 18, 19, 20, 21 e 22.

NEWMAN, M. *Networks: An Introduction*. [S.l.]: OUP Oxford, 2010. ISBN 9870191500701. Citado 10 vezes nas páginas 16, 19, 20, 21, 22, 26, 27, 28, 31 e 33.

LESKOVEC, J. *Dynamics of large networks*. Tese (Doutorado) — Carnegie Mellon University, School of Computer Science, Machine Learning . . . , 2008. Citado 2 vezes nas páginas 18 e 19.

ERDŐS, P.; RÉNYI, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, v. 5, n. 1, p. 17–60, 1960. Citado 2 vezes nas páginas 18 e 50.

GOLBECK, J. *Analyzing the social web*. [S.l.]: Newnes, 2013. Citado 4 vezes nas páginas 19, 20, 21 e 24.

ESTRADA, E. *The structure of complex networks: theory and applications*. [S.l.]: Oxford University Press, 2012. Citado 4 vezes nas páginas 19, 20, 22 e 28.

HANSEN, D.; SHNEIDERMAN, B.; SMITH, M. A. *Analyzing social media networks with NodeXL: Insights from a connected world*. [S.l.]: Morgan Kaufmann, 2010. Citado 4 vezes nas páginas 20, 22, 24 e 25.

WASSERMAN, S.; FAUST, K. *Social network analysis: Methods and applications*. [S.l.]: Cambridge university press, 1994. Citado 4 vezes nas páginas 22, 23, 24 e 42.

NEWMAN, M. *Networks*. [S.l.]: Oxford university press, 2018. Citado 5 vezes nas páginas 22, 23, 24, 25 e 34.

MARTIN, C.; NIEMEYER, P. The role of network size for the robustness of centrality measures. In: SPRINGER. *International Conference on Complex Networks and Their Applications*. [S.l.], 2019. p. 40–51. Citado na página 22.

AL-TAIE, M. Z.; KADRY, S. *Python for graph and network analysis*. [S.l.]: Springer, 2017. Citado na página 23.

BISHT, J. *Closeness Centrality*. 2018. `https://www.geeksforgeeks.org/closeness-centrality-centrality-measure/`. Acessado em: 2020-09-30. Citado 2 vezes nas páginas 23 e 24.

ZAFARANI, R.; ABBASI, M. A.; LIU, H. *Social media mining: an introduction*. [S.l.]: Cambridge University Press, 2014. Citado na página 24.

VARLAMIS, I.; EIRINAKI, M.; LOUTA, M. A study on social network metrics and their application in trust networks. In: IEEE. *2010 International Conference on Advances in Social Networks Analysis and Mining*. [S.l.], 2010. p. 168–175. Citado na página 24.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. *The pagerank citation ranking: Bringing order to the web*. [S.l.], 1999. Citado 2 vezes nas páginas 25 e 26.

ZINOVIEV, D. *Complex network analysis in Python: Recognize-construct-visualize-analyze-interpret*. [S.l.]: Pragmatic Bookshelf, 2018. Citado 2 vezes nas páginas 25 e 26.

ROGERS, I. The google pagerank algorithm and how it works. 2002. Citado na página 25.

345KAI. *PageRanks-Example*. 2007. https://pt.wikipedia.org/wiki/PageRank#/media/Ficheiro:PageRanks-Example.svg. Acessado em: 2020-09-30. Citado 2 vezes nas páginas 25 e 26.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. 1998. Citado na página 26.

SOBOLEVSKY, S.; CAMPARI, R.; BELYI, A.; RATTI, C. General optimization technique for high-quality community detection in complex networks. *Physical Review E*, APS, v. 90, n. 1, p. 012811, 2014. Citado na página 27.

GIRVAN, M.; NEWMAN, M. E. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, National Acad Sciences, v. 99, n. 12, p. 7821–7826, 2002. Citado 4 vezes nas páginas 27, 28, 29 e 30.

LANCICHINETTI, A.; FORTUNATO, S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, APS, v. 80, n. 1, p. 016118, 2009. Citado na página 28.

FORTUNATO, S.; HRIC, D. Community detection in networks: A user guide. *Physics reports*, Elsevier, v. 659, p. 1–44, 2016. Citado 3 vezes nas páginas 28, 33 e 34.

JORA, C.; CHIRA, C. Evolutionary community detection in complex and dynamic networks. In: IEEE. *2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP)*. [S.l.], 2016. p. 127–134. Citado na página 28.

ZHANG, S.; NING, X.-M.; ZHANG, X.-S. Graph kernels, hierarchical clustering, and network community structure: experiments and comparative analysis. *The European Physical Journal B*, Springer, v. 57, n. 1, p. 67–74, 2007. Citado 2 vezes nas páginas 28 e 30.

LANCICHINETTI, A.; FORTUNATO, S. Community detection algorithms: a comparative analysis. *Physical review E*, APS, v. 80, n. 5, p. 056117, 2009. Citado 2 vezes nas páginas 28 e 61.

MALLIAROS, F. D.; VAZIRGIANNIS, M. Clustering and community detection in directed networks: A survey. *Physics Reports*, Elsevier, v. 533, n. 4, p. 95–142, 2013. Citado na página 28.

RAVASZ, E.; SOMERA, A. L.; MONGRU, D. A.; OLTVAI, Z. N.; BARABÁSI, A.-L. Hierarchical organization of modularity in metabolic networks. *science*, American Association for the Advancement of Science, v. 297, n. 5586, p. 1551–1555, 2002. Citado 2 vezes nas páginas 30 e 31.

NEWMAN, M. E.; GIRVAN, M. Finding and evaluating community structure in networks. *Physical review E*, APS, v. 69, n. 2, p. 026113, 2004. Citado 3 vezes nas páginas 30, 31 e 36.

NEWMAN, M. E. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, National Acad Sciences, v. 103, n. 23, p. 8577–8582, 2006. Citado na página 30.

XAVIER, C. R. *Maximização De Influência Em Redes Complexas Para O Modelo De Spreading Activation: Estudo De Caso Em Uma Rede De Telefonia*. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil., 2015. Citado na página 31.

NEWMAN, M. E. J. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, National Academy of Sciences, v. 103, n. 23, p. 8577–8582, 2006. ISSN 0027-8424. Disponível em: <https://www.pnas.org/content/103/23/857>. Citado na página 31.

BARABÁSI, A.-L.; GULBAHCE, N.; LOSCALZO, J. Network medicine: a network-based approach to human disease. *Nature reviews genetics*, Nature Publishing Group, v. 12, n. 1, p. 56–68, 2011. Citado 2 vezes nas páginas 31 e 33.

RAGHAVAN, U. N.; ALBERT, R.; KUMARA, S. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, APS, v. 76, n. 3, p. 036106, 2007. Citado na página 32.

LIU, X.; MURATA, T. Advanced modularity-specialized label propagation algorithm for detecting communities in networks. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 389, n. 7, p. 1493–1500, 2010. Citado na página 32.

COSCIA, M.; ROSSETTI, G.; GIANNOTTI, F.; PEDRESCHI, D. Demon: a local-first discovery method for overlapping communities. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2012. p. 615–623. Citado 3 vezes nas páginas 32, 35 e 55.

XIE, J.; SZYMANSKI, B. K.; LIU, X. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In: IEEE. *2011 ieee 11th international conference on data mining workshops*. [S.l.], 2011. p. 344–349. Citado 3 vezes nas páginas 32, 35 e 55.

PALLA, G.; DERÉNYI, I.; FARKAS, I.; VICSEK, T. Uncovering the overlapping community structure of complex networks in nature and society. *nature*, Nature Publishing Group, v. 435, n. 7043, p. 814, 2005. Citado 3 vezes nas páginas 33, 34 e 35.

NICOSIA, V.; MANGIONI, G.; CARCHIOLO, V.; MALGERI, M. Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment*, IOP Publishing, v. 2009, n. 03, p. P03024, 2009. Citado 3 vezes nas páginas 33, 36 e 56.

ADAMCSEK, B.; PALLA, G.; FARKAS, I.; DERÉNYI, I.; VICSEK, T. Cfinder: Locating cliques and overlapping modules in biological networks. *Bioinformatics*, Oxford University Press, v. 22, n. 8, p. 1021–1023, 2006. Citado na página 35.

DERÉNYI, I.; PALLA, G.; VICSEK, T. Clique percolation in random networks. *Physical review letters*, APS, v. 94, n. 16, p. 160202, 2005. Citado na página 35.

LÁZÁR, A.; ÁBEL, D.; VICSEK, T. Modularity measure of networks with overlapping communities. *EPL (Europhysics Letters)*, IOP Publishing, v. 90, n. 1, p. 18001, 2010. Citado na página 36.

BIKHCHANDANI, S.; HIRSHLEIFER, D.; WELCH, I. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, The University of Chicago Press, v. 100, n. 5, p. 992–1026, 1992. Citado na página 39.

LÓPEZ-PINTADO, D. Diffusion in complex social networks. *Games and Economic Behavior*, Citeseer, v. 62, n. 2, p. 573–590, 2008. Citado 2 vezes nas páginas 39 e 51.

GULATI, R. Network location and learning: The influence of network resources and firm capabilities on alliance formation. *Strategic management journal*, Wiley Online Library, v. 20, n. 5, p. 397–420, 1999. Citado na página 39.

TIDD, J.; BESSANT, J. R. *Managing innovation: integrating technological, market and organizational change*. [S.l.]: John Wiley & Sons, 2018. Citado 2 vezes nas páginas 39 e 42.

KEMPE, D.; KLEINBERG, J.; TARDOS, É. Maximizing the spread of influence through a social network. In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2003. p. 137–146. Citado 4 vezes nas páginas 40, 41, 42 e 59.

ZHANG, H.; MISHRA, S.; THAI, M. T.; WU, J.; WANG, Y. Recent advances in information diffusion and influence maximization in complex social networks. *Opportunistic Mobile Social Networks*, CRC Press, v. 37, n. 1.1, p. 37, 2014. Citado 3 vezes nas páginas 40, 43 e 59.

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*. [S.l.: s.n.], 2001. p. 57–66. Citado 2 vezes nas páginas 40 e 41.

KEMPE, D.; KLEINBERG, J.; TARDOS, E. Maximizing the spread of influence through a social network. *Theory of Computing*, Theory of Computing, v. 11, n. 4, p. 105–147, 2015. Disponível em: <http://www.theoryofcomputing.org/articles/v011a00>. Citado na página 40.

CHEN, W.; WANG, C.; WANG, Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: ACM. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2010. p. 1029–1038. Citado na página 40.

LESKOVEC, J.; KRAUSE, A.; GUESTRIN, C.; FALOUTSOS, C.; VANBRIESEN, J.; GLANCE, N. Cost-effective outbreak detection in networks. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2007. p. 420–429. Citado 2 vezes nas páginas 41 e 42.

CHEN, W.; LAKSHMANAN, L. V.; CASTILLO, C. Information and influence propagation in social networks. *Synthesis Lectures on Data Management*, Morgan & Claypool Publishers, v. 5, n. 4, p. 1–177, 2013. Citado na página 41.

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*. [S.l.: s.n.], 2009. p. 199–208. Citado 3 vezes nas páginas 41, 59 e 86.

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*. [S.l.: s.n.], 2011. p. 47–48. Citado 2 vezes nas páginas 42 e 59.

SHAKARIAN, P.; BHATNAGAR, A.; ALEALI, A.; SHAABANI, E.; GUO, R. *Diffusion in Social Networks*. [S.l.: s.n.], 2015. Citado 4 vezes nas páginas 44, 45, 46 e 47.

GOLDENBERG, J.; LIBAI, B. Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata. v. 9, 01 2001. Citado na página 45.

CRESWELL, J. W.; CRESWELL, J. D. *Research design: Qualitative, quantitative, and mixed methods approaches*. [S.l.]: Sage publications, 2017. Citado na página 48.

LESKOVEC, J.; KREVL, A. *SNAP Datasets: Stanford Large Network Dataset Collection*. jun. 2014. `http://snap.stanford.edu/data`. Citado na página 49.

KUNEGIS, J. Konect: the koblenz network collection. In: *Proceedings of the 22nd International Conference on World Wide Web*. [S.l.: s.n.], 2013. p. 1343–1350. Citado na página 49.

ROBBIN, J. L.; NEWMAN, L. C.; FORSTALL, S. J.; LEMAY, S. O.; GEDIKIAN, S. S.; MULLER, M.; ALSINA, T. *Management of downloads from a network-based digital data repository based on network performance*. [S.l.]: Google Patents, dez. 1 2015. US Patent 9,201,895. Citado 2 vezes nas páginas 49 e 50.

ROSSI, R. A.; AHMED, N. K. The network data repository with interactive graph analytics and visualization. In: *AAAI*. [s.n.], 2015. Disponível em: <http://networkrepository.co>. Citado na página 49.

AYDIN, M. N.; PERDAHCI, N. Z. Analysis and implications of the giant component for an online interactive platform. 2016. Citado na página 50.

SHAKARIAN, P.; BHATNAGAR, A.; ALEALI, A.; SHAABANI, E.; GUO, R. *Diffusion in social networks*. [S.l.]: Springer, 2015. Citado na página 51.

SALAVATI, C.; ABDOLLAHPOURI, A.; MANBARI, Z. Bridgerank: A novel fast centrality measure based on local structure of the network. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 496, p. 635–653, 2018. Citado 2 vezes nas páginas 51 e 55.

LANCICHINETTI, A.; FORTUNATO, S.; RADICCHI, F. Benchmark graphs for testing community detection algorithms. *Physical review E*, APS, v. 78, n. 4, p. 046110, 2008. Citado na página 52.

ROSSETTI, G.; MILLI, L.; CAZABET, R. Cdlib: a python library to extract, compare and evaluate communities from complex networks. *Applied Network Science*, Springer, v. 4, n. 1, p. 52, 2019. Citado na página 56.

ROSSETTI, G.; MILLI, L.; RINZIVILLO, S.; SÎRBU, A.; PEDRESCHI, D.; GIANNOTTI, F. Ndlib: a python library to model and analyze diffusion processes over complex networks. *International Journal of Data Science and Analytics*, Springer, v. 5, n. 1, p. 61–79, 2018. Citado na página 57.

MILLI, L.; ROSSETTI, G.; PEDRESCHI, D.; GIANNOTTI, F. Active and passive diffusion processes in complex networks. *Applied network science*, Springer, v. 3, n. 1, p. 1–15, 2018. Citado na página 59.

VIEIRA, V. d. F.; XAVIER, C. R.; EVSUKOFF, A. G. Comparing the community structure identified by overlapping methods. In: SPRINGER. *International Conference on Complex Networks and Their Applications*. [S.l.], 2019. p. 262–273. Citado na página 85.

ZHOU, X.; LIU, Y.; WANG, J.; LI, C. A density based link clustering algorithm for overlapping community detection in networks. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 486, p. 65–78, 2017. Citado na página 85.

Anexos

ANEXO A – Resultados das Execuções dos Algoritmos de Detecção de Comunidades.

A Tabela abaixo disponibiliza os resultados obtidos dos métodos de detecção de comunidades DEMON e SLPA, executados pelo método *pool_grid_filter*, retornando a melhor partição encontrada de acordo com a medida Q_{ov} (medida de modularidade).

Redes	DEMON			SLPA			OMNI - LFK
	n_c	ϵ	Q_{ov}	n_c	t ; r	Q_{ov}	
GrQc	233	0.25	0.1107	334	18; 0.1	0.1174	0.26427
HepTh	369	0.25	0.10492	681	21; 0.2	0.0975	0.1940
Dolphins	3	0.25	0.1067	2	20; 03	0.1391	0.3227
Football	1	0.25	0.1253	1	10; 0.2	0.1436	0.0
Jazz	4	0.35	0.1336	1	19; 0.4	0.1442	0.3472
Karate	2	0.25	0.1206	1	12; 0.1	0.1442	0.0
Moreno	5	0.25	0.0347	19	18; 0.1	0.0244	0.0441
Poolbooks	3	0.25	0.1067	3	15; 0.2	0.1362	0.1958
Epinions1	284	0.25	0.1156	10963	16; 0.1	0.0235	0.2306
sign-epinions	358	0.25	0.1935	16133	15; 0.2	0.0259	0.1465
wiki-vote	23	0.45	0.1132	245	19; 0.1	0.0252	0.1005
LFR1	32	0.25	0.1064	10	14; 0.1	0.1267	0.1077
LFR2	36	0.25	0.1276	4	14; 0.1	0.1255	0.2663
LFR3	45	0.25	0.0658	7	15; 0.1	0.13	0.0838
LFR5	158	0.25	0.0267	6	20; 0.4	0.1258	0.0516
LFR6	163	0.25	0.0272	19	18; 0.2	0.1218	0.0451
LFR7	111	0.25	0.0134	89	20; 0.3	0.0894	0.0
LFR8	108	0.25	0.0132	66	20; 0.3	0.1019	0.0

Tabela 9: Resultados das execuções de detecção de comunidades para os algoritmos Demon e SLPA das redes de estudo. Para cada rede é mostrado: (n_c), o número de comunidades detectadas pelo algoritmo; (ϵ) parâmetro do método Demon e t e r parâmetros do método SLPA; (Q_{ov}), valor da modularidade para o resultado das partições alcançadas; (OMNI) (*overlapping mutual normalized information*), comparação entre os métodos da qualidade da partição em relação a informação mútua normalizada.