UMA ABORDAGEM PARA DETECÇÃO DE COMUNIDADES A PARTIR DE SEQUÊNCIAS DE INTERAÇÕES SOCIAIS

JEANCARLO CAMPOS LEÃO

UMA ABORDAGEM PARA DETECÇÃO DE COMUNIDADES A PARTIR DE SEQUÊNCIAS DE INTERAÇÕES SOCIAIS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Alberto H. F. Laender
Coorientador: Pedro O. S. Vaz de Melo

Belo Horizonte
Abril de 2018

Leão, Jeancarlo Campos

L437a Uma abordagem para detecção de comunidades a partir de sequências de interações sociais / Jeancarlo Campos Leão. — Belo Horizonte, 2018
xiii, 63 f. : il. ; 29cm

Dissertação (mestrado) – Universidade Federal de Minas Gerais – Departamento de Ciência da Computação.
Orientador: Alberto Henrique Frade Laender.
Coorientador: Pedro Olmo Stancioli Vaz de Melo.

1. Computação - Teses. 2. Redes sociais on-line. 3. Banco de dados temporais. 4. Detecção de comunidades. I. Orientador. II. Coorientador. III. Título.

CDU 519.6*04(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Uma Abordagem para Detecção de Comunidades a partir de Sequências de Interações Sociais

JEANCARLO CAMPOS LEAO

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ALBERTO HENRIQUE FRADE LAENDER - Orientador
Departamento de Ciência da Computação - UFMG

PROF. PEDRO OLMO STANCIOLI VAZ DE MELO - Coorientador
Departamento de Ciência da Computação - UFMG

PROF. FABRÍCIO BENEVENUTO DE SOUZA
Departamento de Ciência da Computação - UFMG

PROFA. MIRELLA MOURA MORO
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 24 de abril de 2018.

Dedico este trabalho a meus filhos Jean J. e Juancarlo.

Agradecimentos

Agradeço primeiramente a Deus por iluminar o meu caminho, permitir minhas realizações e me dar proteção para seguir em frente.

Meus sinceros agradecimentos a todos que me apoiaram para o desenvolvimento deste trabalho. Gostaria de agradecer em especial:

À minha família, pelo suporte e compreensão, pois foram essenciais.

Aos meus melhores amigos, pela companhia e ajuda. Agradeço também aos que, mesmo de longe, compreenderam a necessidade do meu afastamento temporal, mantendo fortes os nossos laços de amizade.

Aos professores Alberto H. F. Laender e Pedro O. S. Vaz de Melo, pela disposição, ensinamento e orientação, fundamentais em todas as etapas deste trabalho.

Aos professores que também foram fonte de motivação, em especial, no início da minha caminhada, à professora Raquel Prates e ao professor Rodrygo Santos e também pela colaboração, à professora Michele Brandão.

Aos colegas do Laboratório de Bancos de Dados e da UFMG, pela amizade e pelo conhecimento construído.

Aos funcionários do Departamento de Ciência da Computação da UFMG pela atenção às minhas dúvidas e solicitações.

Ao Instituto Federal do Norte de Minas Gerais - IFNMG, pela concessão do afastamento e bolsa de qualificação no âmbito do Programa de Bolsas para Qualificação de Servidores.

Ao projeto MASWeb e ao CNPq, FAPEMIG e CAPES, pelo apoio financeiro às atividades oriundas desta dissertação.

"Your focus determines your reality."
(Qui-Gon Jinn)

Resumo

A topologia de uma rede social e o aspecto temporal das interações entre um par de vértices indicam a força do relacionamento entre eles e permitem classificá-lo. Por exemplo, um relacionamento pode ser classificado como persistente e forte com base, respectivamente, na regularidade com que as interações ocorrem e no número de vizinhos em comum do par de vértices envolvido. Por outro lado, um relacionamento raro e fraco é, em geral, aleatório e causa ruído em uma rede social, ocultando a estrutura mais significativa da rede e impedindo uma análise precisa. Nesta dissertação, propomos um arcabouço para preprocessamento de dados de redes sociais que explora propriedades temporais e topológicas de suas sequências de interações reais e sintéticas para melhorar a detecção de comunidades estáticas por algoritmos existentes. Ao remover relacionamentos aleatórios, verificamos por meio de múltiplas fontes de evidência que, em mais de 80% dos casos, as redes sociais convergem para uma topologia com relacionamentos mais puramente sociais e estruturas de comunidade com maior qualidade.

Abstract

The topology of a social network and the temporal aspect of the interactions between a pair of vertices indicate the strength of the relationship between them and allow to classify it. For example, a relationship can be classified as persistent and embedded based, respectively, on the regularity with which interactions occur and on the number of common neighbors of the pair of vertices involved. On the other hand, a rare and less embedded relationship is generally random and represents noise in a social network, hiding the most significant structure of the network and preventing an accurate analysis. In this dissertation, we propose a framework to handle social network data that exploits temporal and topological features of its sequences of real and synthetic interactions to improve the detection of static communities by existing algorithms. By removing random relationships, we verified by means of multiple sources of evidence that in more than 80% of the cases, the social networks converge to a topology with more purely social relationships and higher quality community structures.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A modelagem de uma rede real como um grafo foi primeiramente apresentada no artigo de Leonhard Euler [Euler, 1736] que no século XVII demonstrou não haver solução para o problema de atravessar as sete pontes de Königsberg, sem repetir nenhuma delas. Desde então, diversas áreas do conhecimento científico abordam problemas reais utilizando a teoria de grafos para modelar e propor soluções para diversos tipos de sistema [Barabási & Pósfai, 2016; Barrat et al., 2008; David et al., 2010].

Dentre os diversos tipos de sistema, as redes sociais também podem ser modeladas como um grafo, em que os vértices representam pessoas e as arestas algum tipo de relacionamento entre essas pessoas. Uma aresta pode indicar, por exemplo, se duas pessoas são amigas, se trabalham no mesmo local, se têm interesses em comum ou se colaboraram em um projeto de pesquisa. A definição das arestas é normalmente feita a partir do histórico de interações entre pares de pessoas, que pode ser composto por ligações telefônicas, troca de *e-mails*, coautoria de artigos científicos, encontros, etc. Dessa forma, o histórico dessas interações pode indicar relacionamentos de diferentes intensidades e diferentes naturezas, sendo fundamentais na constituição topológica da rede e na sua evolução temporal [Holme, 2015; Moreno, 1953].

1.1 Motivação

Apesar das redes sociais serem, por definição, representações estáticas, os relacionamentos entre as pessoas que compõem tais redes podem mudar ao longo do tempo, por exemplo, duas pessoas que não são amigas hoje podem ser no futuro. Contudo, pode-se estimar quais são os relacionamentos reais e atuais de uma rede social olhando apenas para seu histórico de interações [David et al., 2010]. Formalmente, as redes sociais podem se alterar estruturalmente se os padrões nas interações entre seus mem-

bros se alterarem ao longo do tempo. Nesse cenário dinâmico, um grafo estático não é adequado para representá-las [De Domenico et al., 2015; He & Chen, 2015; Kostakos, 2009; Orke et al., 2013]. Assim, essas redes dinâmicas podem ser representadas por um grafo temporal ou por uma sequência de grafos, em que os vértices representam as pessoas e as arestas são definidas em função do histórico recente de suas interações.

Sobre representações de redes como essas, estáticas ou dinâmicas, são diversas as aplicações e nesta dissertação abordamos a detecção de comunidades. O problema de detecção de comunidades é amplamente estudado no contexto das redes sociais, por exemplo, para a mitigação de doenças infecciosas em escolas pela identificação dos estudantes e professores que, por algum motivo, se reúnem no mesmo espaço ao mesmo tempo [Gemmetto et al., 2014; Nunes et al., 2017]. A detecção de comunidades de pesquisadores que trabalham em uma mesma área de pesquisa [Alves, 2013] é outro exemplo muito importante, para a comunidade científica e para a sociedade em geral.

A abordagem usual para detectar comunidades em redes sociais funciona da seguinte maneira. Primeiro, um grafo estático representando os relacionamentos sociais entre as pessoas é dado como entrada. Depois, um algoritmo processa a rede e retorna subconjuntos dos vértices, muitas vezes disjuntos, correspondentes às comunidades detectadas na rede. Para o caso em que apenas o histórico de interações é conhecido, a abordagem mais comum é gerar um grafo estático agregado a partir delas e depois executar o processo mencionado acima [Holme & Saramäki, 2012; Lancichinetti et al., 2009; Mucha et al., 2010; Nicosia et al., 2013]. O problema dessa abordagem é que, se os padrões de interação variarem muito ao longo do histórico ou se a função de agregação não for acurada, as comunidades detectadas poderão não refletir a realidade.

Muitos estudos sobre detecção de comunidades usam grafos estáticos devido à maior dificuldade de considerar o aspecto temporal [Greene et al., 2010; Holme, 2015]. No entanto, em sua maioria, os sistemas não são realmente estáticos [Holme, 2015; Orke et al., 2013], o que significa que não considerar o aspecto temporal pode causar perda de informação em relação à ordem e proximidade das interações, ou seja, o padrão de evolução da estrutura da comunidade é perdido [Greene et al., 2010]. Portanto, essa simplificação pode gerar relacionamentos sociais que são deslocados temporalmente, o que pode levar a erros na participação de indivíduos em suas respectivas comunidades.

Por exemplo, considere um grupo de pessoas que não se conhecem e trocam muitos *e-mails* em um único dia, mas depois não se comunicam novamente. Agora, considere um outro grupo de pessoas que trocaram essa mesma quantidade de mensagens, mas de forma regular ao longo do histórico de interações. Se a função de agregação considerar unicamente o total de mensagens trocadas entre duas pessoas, então esses relacionamentos terão a mesma intensidade no grafo estático agregado, embora

eles sejam significativamente diferentes entre si. Apesar dos relacionamentos entre os membros de ambos os grupos terem a mesma topologia quando se considera um grafo estático agregado, a dimensão temporal permite diferenciar esses relacionamentos e, consequentemente, a estrutura da comunidade envolvendo esses grupos de pessoas.

De fato, as redes formadas pela agregação de interações estão sujeitas a uma grande variedade de ruídos. Isso significa que uma aresta no grafo estático agregado pode ser, na verdade, fruto de interações aleatórias entre pares de pessoas. Interações aleatórias referem-se às interações entre pares de indivíduos que muito provavelmente não irão interagir novamente. Este é o caso de *e-mails* enviados para um endereço errado ou quando um contato é adicionado apenas devido às facilidades oferecidas por uma mídia social [Abufouda & Zweig, 2015]. A aleatoriedade também está associada a interações efêmeras como, por exemplo, encontros ocasionais ou coautorias de artigos entre pesquisadores com interesses distantes.

Um grande desafio do problema de detecção de comunidades está na dificuldade de avaliar os métodos propostos [Almeida et al., 2011, 2012; Fortunato, 2010; Yang & Leskovec, 2015]. Parte dessa dificuldade reside no fato de que ainda não existe uma definição universalmente aceita para o conceito de comunidade [Abrahao et al., 2012; Fortunato, 2010; Palla et al., 2005] e para a qualidade de uma comunidade [Hric et al., 2014; Rossetti & Cazabet, 2017]. Por outro lado, considerando comunidades reais, uma propriedade fundamental compartilhada por diferentes definições é a presença de relacionamentos sociais mais fortes dentro da comunidade do que fora dela e que geralmente se mantêm ao longo do tempo [David et al., 2010; He & Chen, 2015; Kivelä et al., 2014; Kossinets & Watts, 2006]. Isso motiva o uso da dimensão temporal para complementar o conjunto de propriedades que permitem avaliar os relacionamentos. Além disso, as propriedades topológicas e temporais são universais, ou seja, são independentes de outras propriedades específicas de cada domínio.

Assim, a ideia principal desta dissertação é analisar o problema de detecção de comunidades a partir de sequências de interações. Nesse contexto, nossa hipótese principal é que o grafo estático agregado, quando construído sem cuidado a partir dessas interações, contém diversos relacionamentos que são fruto de interações aleatórias. Nesses casos, os métodos de detecção de comunidades podem gerar comunidades com diversos relacionamentos fracos e raros entre os seus membros. Por isso, avaliamos e diferenciamos relacionamentos que normalmente seriam considerados iguais em uma rede, apoiando-nos em teorias como a força dos laços [Granovetter, 1973], que há décadas foi consolidada em diversos campos da ciência. Portanto, avaliar quão real é um relacionamento social é inevitável para obter uma representação de alta qualidade da estrutura de comunidade presente no sistema estudado [Abufouda & Zweig, 2017].

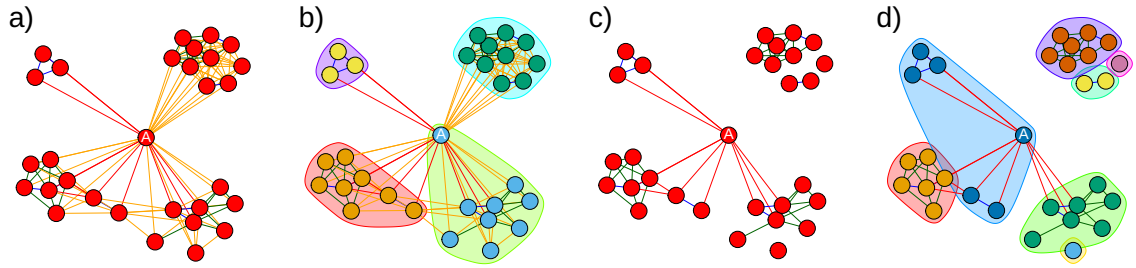

Figura 1.1: Exemplo de como o ruído causado por arestas aleatórias pode afetar a detecção de comunidades em redes sociais. a) Uma rede formada a partir das colaborações científicas de F. M. Peeters da *American Physical Society*. Os vértices são pesquisadores e há uma aresta entre dois vértices se os respectivos pesquisadores aparecem como coautores de um artigo no passado. b) Comunidades detectadas pelo algoritmo *Louvain* [Blondel et al., 2008] na rede retratada em a). c) A mesma rede depois da remoção de arestas aleatórias. d) Comunidades detectadas pelo algoritmo *Louvain* na rede filtrada, que são diferentes das descritas em b).

A Figura 1.1 ilustra o papel da dimensão temporal no processo de detecção de comunidades estáticas. Especificamente, a Figura 1.1a mostra a rede de ego do pesquisador F. M. Peeters, formada a partir de suas interações históricas de coautoria. Nessa rede, um vértice representa um pesquisador e há uma aresta entre dois vértices se os respectivos pesquisadores publicaram um artigo juntos no passado. Já a Figura 1.1b mostra as comunidades detectadas nessa rede pelo algoritmo *Louvain* [Blondel et al., 2008]. A Figura 1.1c, por sua vez, mostra a mesma rede depois de removidas as arestas provenientes de interações aleatórias usando-se o algoritmo *RECAST* [Vaz de Melo et al., 2015]. Observe na Figura 1.1d que, quando o algoritmo *Louvain* é aplicado à rede, a estrutura de comunidades revelada é surpreendentemente diferente da descrita na Figura 1.1b. Mais importante, mostramos mais adiante nesta dissertação que a estrutura de comunidades detectada na rede filtrada é mais representativa dos relacionamentos sociais subjacentes do que a estrutura detectada sem a etapa de filtragem. Neste contexto, buscamos responder as seguintes questões de pesquisa:

Q1. Qual é o efeito da filtragem de relacionamentos aleatórios para algoritmos de detecção de comunidades?

Q2. Como avaliar a qualidade das comunidades geradas quando a detecção é feita a partir de uma rede filtrada?

Q3. A filtragem de relacionamentos aleatórios é sempre benéfica ou varia com o algoritmo de detecção utilizado e com o tipo de interação existente?

1.2 Objetivos e Contribuições

O principal objetivo desta dissertação é o desenvolvimento de um arcabouço para melhorar a detecção de comunidades a partir de sequências de interações sociais. O arcabouço tem como princípio a eliminação de arestas aleatórias do grafo agregado com base na análise do histórico de interações. Esse histórico de interações deverá conter os pares de vértices que representam pessoas que interagiram e os instantes de tempo em que essas interações ocorreram.

Para avaliar a melhoria obtida pelo nosso arcabouço, utilizamos múltiplas abordagens de avaliação. Especificamente, avaliamos a qualidade das comunidades obtidas considerando o consenso entre diferentes linhas de evidência. Essa avaliação foi feita utilizando-se dez redes e seis algoritmos de detecção de comunidades considerados o estado da arte, e comparando os resultados obtidos com os de outros métodos de remoção de ruído.

Neste contexto, as principais contribuições desta dissertação são:

- Um arcabouço para filtragem de ruídos em redes sociais que permite melhorar a detecção de comunidades estáticas por algoritmos existentes e que possibilita o uso de diferentes modelos de força dos relacionamentos para distinguir os relacionamentos sociais dos aleatórios [Leão et al., 2017a,b].

- Uma abrangente avaliação do arcabouço proposto que se baseia na análise de diferentes evidências que indicam a melhoria da qualidade das comunidades detectadas [Leão et al., 2018]. Mais especificamente, determinamos que a filtragem da rede melhora a detecção de comunidades considerando três estratégias de avaliação: estrutural, funcional e comparação com *baselines*.

1.3 Organização da Dissertação

O restante desta dissertação está organizado da seguinte forma:

- O Capítulo 2 apresenta uma breve revisão dos principais conceitos sobre grafos e redes sociais, de modelagem para representação de redes sociais e das principais métricas utilizadas para avaliação da estrutura das comunidades.

- O Capítulo 3 apresenta o arcabouço proposto para melhorar a detecção de comunidades, as técnicas de detecção de comunidades que foram utilizadas em nossos experimentos, e as etapas e estratégias de avaliação do nosso arcabouço.

- O Capítulo 4 apresenta uma análise detalhada dos resultados experimentais obtidos com a aplicação do arcabouço proposto em redes reais e simuladas. Primeiramente, a Seção 4.1 descreve os dados utilizados nos experimentos e, em seguida, a Seção 4.2 detalha as configurações do arcabouço. Concluindo o capítulo, a Seção 4.3 apresenta uma análise das evidências obtidas que demonstram a melhoria da detecção de comunidades pela filtragem de relacionamentos aleatórios.

- Finalmente, o Capítulo 5 apresenta as nossas conclusões e algumas considerações sobre trabalhos futuros.

Capítulo 2

Fundamentos e Trabalhos Relacionados

2.1 Conceitos Básicos sobre Grafos

De acordo com a teoria dos grafos, um grafo não direcionado G é definido por um par $G = (V, E)$, em que $V = \{v_1, v_2, \ldots, v_n\}$ é um conjunto não vazio de elementos chamados vértices ou nodos e $E = \{e_1, e_2, \ldots, e_m\}$ é um conjunto de pares não ordenados de diferentes vértices chamados arestas. Assim, dada uma aresta $e_l = (v_i, v_j)$ unindo dois vértices v_i e v_j, esses vértices são ditos conectados. Em outras palavras, podemos afirmar que a existência de uma aresta entre v_i e v_j significa que eles são vizinhos. O número total de arestas em um grafo determina o seu tamanho e o número de vértices determina a sua ordem.

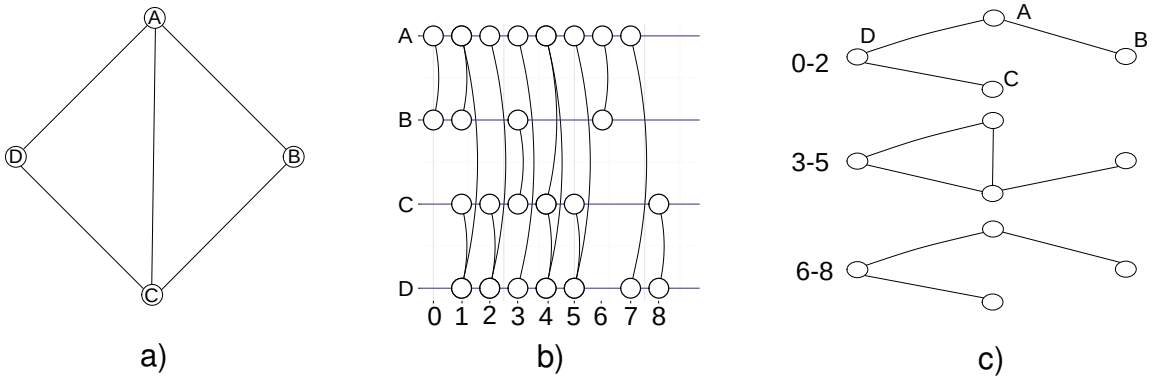

Figura 2.1: Diferentes representações de uma mesma rede social: a) grafo simples ou estático; b) sequência de interações temporais; c) grafo de agregação temporal

Quando um grafo possui apenas uma aresta entre cada par de vértices, ele é chamado simples ou estático (Figura 2.1a). Sem essa restrição, ou seja, quando o mesmo par de vértices é conectado por múltiplas arestas, esse grafo corresponde a um multigrafo. Em ambos os casos, cada aresta pode possuir propriedades, ou seja, valores associados a elas como uma data, um peso ou um nome.

Um subgrafo G' de um grafo G é um grafo cujo conjunto de vértices é um subconjunto dos vértices de G e cujo conjunto de arestas é um subconjunto das arestas de G. Assim, temos $G' = (V', E')$ tal que $V' \subseteq V$ e $E' \subseteq E$. Um subgrafo induzido por vértices $G[V']$ é um grafo formado por um subconjunto V' de vértices de um grafo G com todas as arestas que ligam pares de vértices em V'. De forma análoga, um subgrafo induzido por arestas $G[E']$ é um grafo formado por um um subconjunto de arestas E' de um grafo G contendo todos os vértices que estão em suas extremidades.

Grau. O grau d_i ou $deg(v_i)$ é um atributo local de um vértice $v_i \in V$ que é definido como o número de arestas incidentes em v_i. O grau médio é uma propriedade global de um grafo, obtida pelo cálculo da média de valores de deg sobre todos os seus vértices.

Coeficiente de agrupamento global. O coeficiente de agrupamento global de um grafo indica o quanto os seus vértices tendem a agrupar-se. É obtido pela razão entre o número de triângulos (trios de vértices interligados por três arestas) e o número de trios abertos e fechados (trio de vértices ligados por duas ou três arestas). Este coeficiente também é conhecido na análise de redes sociais como transitividade [Wasserman & Faust, 1994].

Sobreposição de vizinhança. Também conhecida como Índice de Jaccard [Papadimitriou et al., 2010] ou Similaridade [Vaz de Melo et al., 2015], a sobreposição de vizinhança (*Neighborhood Overlap - NO*) de uma aresta $e = (v_i, v_j)$ é a razão entre o número de vizinhos que são comuns a ambos os vértices v_i e v_j, e o número de vizinhos de pelo menos um dos vértices [David et al., 2010] (Equação 2.1).

$$NO(v_i, v_j) = \frac{|N_{v_i} \cap N_{v_j}|}{|N_{v_i} \cup N_{v_j}|} \tag{2.1}$$

2.2 Redes Sociais

Em termos gerais, uma rede social é um sistema que pode ser matematicamente representado por um grafo. Além disso, um grafo estático é o meio natural para representar uma rede estática e um grafo temporal (variável no tempo ou em evolução) é o meio natural para representar redes altamente dinâmicas [Casteigts et al., 2011; Holme & Saramäki, 2013]. Com base nisso, distinguimos a seguir as redes de relacionamento

das redes de interações. O tamanho de uma rede corresponde à ordem do grafo que a representa, ou seja, o número de entidades representadas, que é denotado por $|V|$.

Sequência de Interações. Em um cenário mais dinâmico, as entidades de uma rede interagem umas com as outras (por exemplo, trocas de *e-mails*) ao longo do tempo (Figura 2.1b). Tais interações podem ser representadas por meio de uma sequência ordenada de arestas $E = \{e_1, ..., e_m\}$. A k-ésima interação é uma tupla $e_k = (\tau, v_i, v_j)$, onde τ é uma propriedade de cada interação que indica o tempo em que a interação ocorreu, e v_i e v_j são as entidades que interagiram uma com a outra.

Agregação em relacionamentos. Muitas vezes, pode ser necessário representar o conjunto de interações entre cada par de indivíduos através de um relacionamento entre eles. Tal representação forma a rede de relacionamentos $G_{os} = (V, R)$ que corresponde à representação estática e geralmente a mais recente de uma rede social[1]. A forma usual de fazer esse mapeamento é através da agregação ao longo de todo o tempo, ou seja, todas as interações entre cada par de vértices são "achatadas" e, assim, representadas por uma única aresta [Holme & Saramäki, 2013; Nicosia et al., 2013]. Portanto, existirá uma e somente uma aresta (relacionamento) do conjunto $R = \{r_1, r_2, ..., r_q\}$ entre cada par de vértices $(v_i, v_j) \in G_{os}$ se existir pelo menos uma aresta (interação) entre os respectivos pares de vértices $(v_i, v_j) \in G$.

Uma rede de relacionamentos contém rótulos como propriedades de suas arestas que identificam o tipo de relacionamento entre seus vértices [Vaz de Melo et al., 2015]. Dado um conjunto de rótulos $L = \{l_1, ..., l_u\}$ e um grafo simples $G_c = (V_c, R_c)$ com entidades $V_c \subseteq V$ e arestas rotuladas $R_c = \{r_1, ..., r_s\}$ que representam relacionamentos, o k-ésimo relacionamento é uma tupla $r_k = (l_\ell, v_i, v_j)$, onde $l_\ell \in L$ é o rótulo que identifica a classe deste relacionamento entre as entidades v_i e v_j.

As interações e os relacionamentos são mencionados ao longo de todo o texto. Assim, o conceito de aresta, quando possível, será acompanhado ou substituído pelo conceito que ela representa, ou seja, um relacionamento em uma rede estática ou uma interação em uma rede temporal.

Rede Temporal Agregada em Janelas de Tempo. É possível construir uma *rede de agregação temporal* G_{ot} a partir de uma sequência de interações. Especificamente, cada grafo $G_\kappa(V_\kappa, E_\kappa)$ em G_{ot} representa a agregação de interações em períodos discretos de tempo κ. Assim, para um dado valor de κ, V_κ inclui todos os vértices que interagiram no κ-ésimo período[2]. Analogamente, as arestas do conjunto E_κ represen-

[1] Considerando que relacionamentos naturalmente evoluem ao longo do tempo [Holme & Saramäki, 2013; Vaz de Melo et al., 2015], nesta dissertação adotamos a sua representação estática e mais recente.

[2] A representação de uma rede social, por exemplo, para modelos de mobilidade humana pode ser feita através de uma rede de agregação temporal. Para isso, a mobilidade é rastreada e representada

tam o emparelhamento de interações entre os pares de vértices (v_i, v_j) que ocorreram durante o período de tempo κ (Figura 2.1c).

Detecção de comunidades. Dado um grafo $G_{os} = (V, R)$, uma comunidade é um subconjunto não vazio $c \subset V$. O problema de detecção de comunidades consiste em encontrar o conjunto de comunidades não sobrepostas $C(G_{os}) = \{c_1, c_2, ..., c_k\}$ no qual cada vértice $v_i \in V$ é associado com uma única comunidade $c_j \in C$. Essa é a definição utilizada por grande parte das técnicas de detecção de comunidades [He & Chen, 2015], de modo que nesta dissertação propomos melhorar a qualidade das comunidades detectadas por essas técnicas.

2.3 Trabalhos Relacionados

O objetivo desta seção é descrever trabalhos existentes relacionados à detecção de comunidades, à força dos laços e à remoção de ruído, bem como apresentar uma visão geral das estratégias existentes comumente usadas para avaliar a eficácia dos algoritmos de detecção de comunidade.

2.3.1 Detecção de Comunidades

Em redes complexas, uma comunidade pode ser vista como um grupo de vértices densamente interligados, mas que são escassamente conectados com o resto da rede [Newman, 2004; Yang & Leskovec, 2015]. Esta não é a única definição de comunidade. Por exemplo, Wang & Hopcroft [2010] caracterizam comunidade como um conjunto de entidades que, além de estarem mais conectadas do que o esperado[3], também podem estar bem conectadas ao resto da rede. Outras características também definem a estrutura de comunidade além da conectividade, como a existência de hierarquia ou de sobreposição entre comunidades [Palla et al., 2005], a dinâmica ou a evolução temporal nessa estrutura [Peixoto & Rosvall, 2017], sua subdivisão estrutural (como em núcleo e em periferia [Leskovec et al., 2008; Wang & Hopcroft, 2010]) ou sua constituição por múltiplos tipos de relacionamentos [Kivelä et al., 2014].

Yang & Leskovec [2015] distinguem a definição estrutural de comunidade da definição funcional. Eles caracterizam a definição estrutural com base em padrões de conectividade, como a densidade de conexões entre os membros da comunidade. Assim, algumas definições estruturais de comunidade levam em consideração apenas a

por uma sequência de interações de contato que então será convertida para a rede temporal agregada.

[3]De acordo com Wang & Hopcroft [2010], uma comunidade é um subconjunto densamente conectado no qual a probabilidade de existir uma aresta entre dois vértices escolhidos aleatoriamente é acima da média.

conectividade interna enquanto outras podem considerar também a conectividade externa dos vértices. Os autores também exemplificam a definição estrutural baseada na métrica de modularidade. Por sua vez, a definição funcional é aquela em que os membros de uma comunidade compartilham funções, propriedades ou propósitos comuns. Por exemplo, os vértices que pertencem a uma mesma comunidade podem possuir os mesmos valores relativos a algum atributo relevante para o domínio da rede, tal como localidade, área de pesquisa, idade ou turma. Além disso, diversos estudos [Fortunato, 2010; Peel et al., 2017; Yang & Leskovec, 2015] demonstram que comunidades funcionais podem ser utilizadas como *ground truth* para avaliar comunidades estruturais.

Baseadas em abordagens distintas, muitas outras técnicas têm sido usadas para a detecção de comunidades em redes estáticas. Dentre as mais populares, podemos citar aquelas propostas por Blondel et al. [2008], Raghavan et al. [2007], Newman & Girvan [2004] e Pons & Latapy [2005] que podem extrair diferentes comunidades de uma mesma rede. Por exemplo, a abordagem que Newman & Girvan [2004] utilizam em seu algoritmo parte da definição que uma comunidade corresponde a cada grupo de vértices que se mantêm conectados após sucessivas remoções de arestas que são mais prováveis de estarem entre esses grupos. Além dessa abordagem, diversas outras que fazem uso de definições distintas de comunidade são relatadas por Coscia et al. [2011].

Diante dessa diversidade de abordagens implementadas em muitos algoritmos para detectar comunidades em redes estáticas, notamos que é grande o número de definições distintas, pois cada algoritmo concentra-se em algumas propriedades específicas das redes e estabelece, explícita ou implicitamente, sua própria definição de comunidade [Coscia et al., 2011; Fortunato, 2010]. Note que isso permite convergir para as observações feitas por Palla et al. [2005], Fortunato [2010] e Abrahao et al. [2012], de que a estrutura de uma comunidade é difícil de definir, quantificar e extrair porque não existe uma definição universalmente aceita. Ademais, em trabalhos distintos como esses, o conceito de comunidade tem sido definido com base mais em características da abordagem utilizada para a sua detecção do que em características da rede.

Para lidar com a diversidade de definições de comunidade, alguns estudos avaliam de forma ampla a sua detecção. Abrahao et al. [2012] e Xie et al. [2013] apresentam análises abrangentes de propriedades de comunidades detectadas por diferentes algoritmos. Eles mostram que as comunidades detectadas e suas propriedades variam consistentemente dentre os algoritmos. Nesses trabalhos e em muitos outros sobre extração de comunidades, apenas os relacionamentos estáticos de mesmo tipo são analisados [Coscia et al., 2011; Fortunato & Hric, 2016; Leskovec et al., 2007; Yang et al., 2016].

Contudo, redes sociais descrevem uma grande variedade de sistemas reais que variam no tempo e evoluem sua estrutura através de interações sucessivas entre entida-

des que surgem ou são removidas com o tempo. Alguns trabalhos abordam o aspecto temporal em redes *Wi-Fi* móveis [Vaz de Melo et al., 2015], em que os vértices representam usuários e as arestas representam pares de usuários que compartilham o mesmo ponto de acesso ao longo do tempo. Da mesma forma, outros trabalhos abordam redes de colaboração científica, em que os vértices representam pesquisadores e as arestas modelam suas interações temporais de coautoria [Alves, 2013; Barabâsi et al., 2002].

2.3.2 Detecção de Comunidades Temporais

Para a detecção de comunidades, alguns algoritmos são aplicados em redes temporais, que representam instantâneos da rede (janelas de tempo) como uma sequência de grafos estáticos. Neste caso, as abordagens usuais detectam comunidades em cada instantâneo de forma independente [Palla et al., 2007] ou iterativamente [Lancichinetti et al., 2009]. Outros algoritmos consideram o aspecto temporal para identificar comunidades dinâmicas, detectando-as globalmente em todos os instantâneos [Cazabet et al., 2010; Mucha et al., 2010].

Infelizmente, as abordagens de detecção de comunidades que exploram aspectos temporais ainda compreendem uma pequena parte das propostas existentes quando comparadas às abordagens para redes estáticas [He & Chen, 2015; Holme, 2015; Leskovec et al., 2007]. Além disso, muitas dessas abordagens para redes temporais se baseiam em estratégias que reutilizam ou adaptam algumas das abordagens para redes estáticas [Cazabet & Amblard, 2014; Yu et al., 2010]. Por esta razão, nesta dissertação escolhemos para estudo o subconjunto de técnicas de detecção de comunidades desenvolvidas para redes estáticas.

2.3.3 Força dos Laços e o Aspecto Temporal

Estudos sociológicos revelam que a topologia em que pares de indivíduos estão envolvidos, ou seja, a estrutura formada pelos seus relacionamentos, indica a a força dos laços entre esses indivíduos, além de evidenciar a formação de comunidades [David et al., 2010]. Em um trabalho seminal, Granovetter [1973] introduz a noção de força dos laços em redes sociais e relata como diferentes classes de relacionamento (fracos e fortes) afetam indivíduos e organizações. A força dos laços alinha-se com a ideia de que os relacionamentos mais fortes ocorrem durante um longo período de tempo entre pessoas cujos círculos sociais se sobrepõem fortemente com o seu próprio [Burt, 1992; Granovetter, 1973], o que geralmente produz uma alta frequência de interação [David et al., 2010]. David et al. [2010] apresentam também uma visão geral da teoria

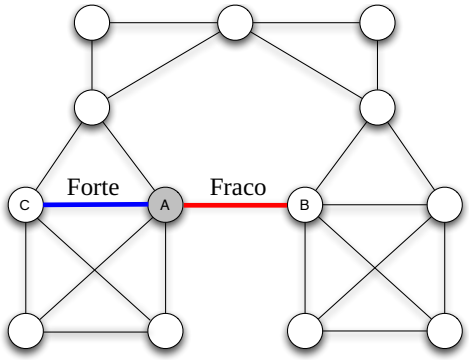

Figura 2.2: Força dos laços modelada pela topologia de rede social. Exemplos de laço forte e de laço fraco destacados para o vértice A. Fonte (adaptado): David et al. [2010].

que relaciona a força dos laços à formação de comunidades. Em conformidade com os estudos de Granovetter, eles revelam que laços fracos são de longa extensão e servem para combinar diferentes comunidades compostas por laços mais fortes, que são aqueles entre vértices que possuem maior número de vizinhos em comum (Figura 2.2).

Além disso, a força de um laço particular pode variar com a evolução de uma rede social ao longo do tempo, quando surgem cenários mais complexos, impulsionados pela combinação de atividades compartilhadas e afiliações de seus membros [Teixeira et al., 2015]. Essa dinâmica também é guiada pela semelhança entre atributos dos indivíduos [Lazarsfeld et al., 1954] e pelo fechamento triádico [Kossinets & Watts, 2006].

A dimensão temporal é considerada no trabalho de Kossinets & Watts [2006] para análise empírica de uma rede social construída a partir de *e-mails* trocados entre estudantes, professores e funcionários de uma faculdade. Esse estudo esclarece a importância da dimensão temporal para identificar e distinguir os fatores que levam ao surgimento, evolução e força dos laços. Outros estudos consideram a agregação temporal para analisar a evolução das redes de colaboração [Alves et al., 2013; Barabâsi et al., 2002]. Além disso, tal conceito é utilizado por Vaz de Melo et al. [2015] para classificar relacionamentos com base em propriedades temporais em redes de mobilidade.

Ademais, é importante lembrar que relacionamentos sociais são formados pela dinâmica das interações. Portanto, considerar o aspecto temporal no estudo de interações sociais torna-se fundamental, pois revela propriedades e padrões que não podem ser percebidos sem considerar tal característica [Barabási & Pósfai, 2016; Holme & Saramäki, 2012; Leão et al., 2017a].

Na prática, métricas simples baseadas nas ideias mencionadas, como a frequência de interação [David et al., 2010] e o número de vizinhos em comum [Shi et al., 2007], são comumente usadas para medir a força dos laços. No entanto, dependendo do contexto,

outros fatores podem ser usados para modelá-la [Adamic & Adar, 2003; Brandão & Moro, 2017; Gilbert & Karahalios, 2009; Marsden & Campbell, 1984].

Em seu trabalho, Adamic & Adar [2003] consideram informações como listas de *e-mails* e o uso de frases comuns em páginas da Web pessoais para medir a semelhança entre os usuários (ou seja, a força dos seus laços). Aral & Walker [2014] analisam a comunicação por troca de mensagens entre pares em uma rede de amizade e identificam o efeito de um modelo de força dos laços, que considera o número de amigos em comum, no aumento da influência social. Brandão & Moro [2017] medem a força de laços de coautoria com base nas propriedades topológicas de redes acadêmicas. Gilbert & Karahalios [2009] propõem uma métrica de força dos laços baseada em informações específicas extraídas do *Facebook*, como o número de dias desde a última comunicação e o número de palavras trocadas em postagens. Abufouda & Zweig [2017] utilizam classificadores baseados em aprendizado de máquina para avaliar a força dos relacionamentos em redes sociais usando apenas a estrutura formada por interações de tipos distintos. Os autores utilizam seu arcabouço para identificar ruído, representado por arestas aleatórias injetadas nas redes. Finalmente, Marsden & Campbell [1984] usam dados socioeconômicos e demográficos para validar a qualidade de uma medida de força dos laços por diferentes aspectos. Embora os autores utilizem um modelo de predição, eles verificam que o aspecto temporal influencia na medida de força dos laços.

É notável que a força dos laços é usada para estimar a probabilidade de pares interagirem novamente. Assim, a probabilidade empírica de formação de laços aumenta com o número de conhecidos mútuos [Kossinets & Watts, 2006]. David et al. [2010] complementam que essa probabilidade é maior para os laços formados por um número maior de interações. Considerando apenas essas dimensões, laços embutidos e formados por interações regulares ou frequentes ocorrem entre vértices com maior probabilidade de interagirem novamente e, por isso, constituem relacionamentos sociais. Por outro lado, laços entre pares de indivíduos com pouca probabilidade de interagirem novamente são considerados relacionamentos aleatórios [Vaz de Melo et al., 2015].

Além dos aspectos topológicos e temporais, outros fatores foram usados para modelar a força dos laços [Gilbert & Karahalios, 2009; Marsden & Campbell, 1984]. Segundo Granovetter [1973], a quantidade de tempo, a intimidade, a intensidade e os serviços recíprocos são propriedades que podem ser exploradas para medir a força dos laços. Para o domínio específico das mídias sociais, propriedades adicionais como as estruturais, as de suporte emocional e a de distância social também são exploradas por Gilbert & Karahalios [2009] para predizer a força dos laços.

Contudo, os atributos necessários para aplicar um modelo nem sempre estão presentes. Isso ocorre em algumas redes sociais porque os atributos de laços com

informações secretas ou sensíveis não podem ser utilizados para esse fim [Shi et al., 2007], ou também porque a existência de alguns atributos é limitada, por exemplo, a um domínio específico de redes sociais. Além disso, a alta dimensionalidade transforma informações excessivamente ricas em dados que são algoritmicamente mais caros para serem coletados, extraídos e utilizados. Assim, a modelagem da força dos laços a partir de características comuns às redes, como a temporal e a topológica, permitem ampliar a diversidade de domínios em que o modelo pode ser aplicado [Gilbert & Karahalios, 2009; Shi et al., 2007].

Em alguns casos, não é possível fazer uso da dimensão temporal porque ela não está disponível no esquema de representação dos dados utilizado, o que torna necessário escolher uma estratégia mais adequada para modelagem dos dados extraídos da rede social, com menor perda de informação e redução do ruído [Barabási & Pósfai, 2016; He & Chen, 2015; Holme & Saramäki, 2012].

De fato, Holme & Saramäki [2012] mostram como o aspecto temporal pode ser representado em redes estáticas e demonstram as implicações desse mapeamento que, por um lado, facilita a análise mas, por outro, resulta em perda de informação. Neste contexto, Rocha et al. [2017] exemplificam em diferentes domínios a importância de considerar o aspecto temporal no estudo de relacionamentos sociais para revelar propriedades e padrões que não podem ser percebidos quando suas interações são totalmente agregadas em redes estáticas. A partir dessas afirmações, nota-se a importância de filtrar relacionamentos sociais quando a dimensão temporal estiver disponível nos dados. Finalmente, as interações podem ser agregadas em redes estáticas, sem o ruído causado por relacionamentos aleatórios e com menor perda informacional na remoção da dimensão temporal.

2.3.4 Remoção de Ruído em Redes Sociais

Normalmente, redes sociais reais contêm ruídos, ou seja, laços que não refletem um relacionamento real e, possivelmente, entidades que possuem apenas laços ruidosos. Esses laços ruidosos (especialmente, falsos positivos) alteram a estrutura real de uma rede e impedem sua análise precisa [Abufouda & Zweig, 2017]. Neste contexto, a avaliação de *links* (ou relacionamentos) é o processo usado para identificar laços ruidosos e não ruidosos, permitindo inferir automaticamente as conexões do mundo real em uma rede [Adamic & Adar, 2003]. Além disso, esse processo possibilita aplicações potenciais em várias situações, como descobrir, rotular e caracterizar comunidades [Abufouda & Zweig, 2015]. Por exemplo, Shi et al. [2007] examinam o efeito da remoção de todos os laços não-transitivos de dois conjuntos de dados de redes sociais reais. Similar-

mente, Ouyang et al. [2016] filtram diversas redes para melhorar a precisão na previsão de *links*, enquanto Spitz et al. [2016] usam métricas distintas para avaliar *links* e identificar interações aleatórias sobre redes biológicas estáticas.

Diferentemente, Klymko et al. [2014] propõem uma abordagem para redes estáticas e direcionadas que consiste em ponderar arestas com base no número de trios de vértices envolvidos. Essa abordagem gera uma rede não direcionada que pode ser usada por algoritmos de detecção de comunidades que fazem proveito de arestas ponderadas.

De modo geral, os métodos propostos para remoção de ruído que visam melhorar a qualidade das comunidades detectadas enfrentam as seguintes limitações: (i) confiam em uma única definição de comunidade, geralmente determinada pelo particionamento de seus vértices orientado à maximização da modularidade; (ii) consideram apenas poucas e pequenas redes estáticas como entrada ou saída. Além disso, esses trabalhos avaliam a melhoria na estrutura das comunidades com base em um reduzido e pouco diversificado conjunto de técnicas para detecção de comunidades, o que contribui para uma interpretação tendenciosa dos resultados.

Com base no melhor do nosso conhecimento, o trabalho de Wen et al. [2011] é o que mais se aproxima do nosso. Neste caso, os autores usam um conceito distinto de ruído que se baseia na presença de "violadores" da estrutura da comunidade, isto é, vértices que são caracterizados por alta centralidade de grau. Essa definição deixa claro que em tal abordagem eles não filtram nenhum ruído causado por arestas que conectam pares de vértices que não são "violadores". Essa abordagem desconsidera ainda a possibilidade de decidir uma comunidade para tais vértices, por exemplo, pela distinção da força de seus relacionamentos. Além disso, em sua análise os autores não apresentam uma interpretação do que causa a violação da estrutura da comunidade por tais vértices nem uma justificativa para o critério de remoção de vértices adotado. Ademais, eles usam apenas o conceito de modularidade para avaliar a eficácia do seu filtro em comunidades que foram detectadas em duas únicas redes estáticas por dois algoritmos distintos. Assim, analisando os principais estudos sobre filtragem de redes, não identificamos abordagens que utilizam o aspecto temporal a partir de uma sequência de interações, para quantificar e reduzir o efeito de relacionamentos aleatórios, especialmente para melhorar a estrutura de comunidade em grandes redes reais de diferentes domínios.

2.3.5 Avaliação da Estrutura de Comunidade

Na maioria dos casos, cada algoritmo possui sua definição de comunidade [Cazabet et al., 2010], ou seja, a abordagem utilizada pelo algoritmo determina o que é uma

comunidade, sem uma predefinição [Coscia et al., 2011; Fortunato, 2010]. Além disso, esses algoritmos são sensíveis a diferentes estruturas de comunidade, topologias ou instâncias de uma rede [Xie et al., 2013]. Redes de domínios distintos, por sua vez, possuem fortes diferenças em sua estrutura, o que significa que a qualidade de seus resultados é variável. Diante disso, torna-se necessário o uso de múltiplas estratégias de avaliação e uma experimentação abrangente com vários algoritmos de detecção de comunidades aplicados a redes sociais de diferentes domínios.

Alguns estudos avaliam a qualidade de uma comunidade através de diferentes linhas de evidência, como as baseadas nas suas características estruturais [Newman, 2004], nas propriedades compartilhadas pelos seus membros [Hric et al., 2016; Yang et al., 2016] ou na comparação com um referencial de qualidade, por exemplo, com uma ou mais técnicas do estado da arte [Yang & Leskovec, 2015]. Nesse sentido, cada linha de evidência pode ser capturada por um conjunto de métricas ou estratégias de avaliação com pressuposto único sobre a qualidade de uma comunidade. Dentre outras estratégias, algumas das mais comuns são exemplificadas a seguir.

2.3.5.1 Avaliação Estrutural

Existem diferentes métricas de qualidade estrutural de comunidades. Yang & Leskovec [2015] exemplificam métricas baseadas em padrões de conectividade, como a densidade de conexões entre os membros da comunidade. Em seu trabalho, os autores avaliam a correlação entre diferentes métricas e mostram que, duas delas, a modularidade e a condutância, não são tão bem correlacionadas. Assim, utilizamos essas métricas neste trabalho, pois essa divergência e sua ampla utilização na avaliação de algoritmos de detecção de comunidades permitem capturar bem diferentes aspectos da estrutura de comunidade. Além disso, é comum que em redes distintas, as características estruturais sejam melhor capturadas por métrica diferentes [Yang & Leskovec, 2015].

Modularidade. Certamente, a métrica mais utilizada para avaliar algoritmos de detecção é a modularidade [Fortunato, 2010; Lambiotte et al., 2008; Newman, 2006a; Newman & Girvan, 2004; Orke et al., 2013; Radicchi et al., 2004; Raghavan et al., 2007; Sah et al., 2014; Šubelj & Bajec, 2011; Wang et al., 2015; Yang et al., 2016]. Além disso, muitos algoritmos que detectam comunidades usam a modularidade como uma métrica a ser maximizada no processo de detecção [Barber & Clark, 2009; Blondel et al., 2008; Clauset et al., 2004; Liu & Murata, 2010; Newman, 2004; Raghavan et al., 2007; Schuetz & Caflisch, 2008]. No entanto, a interpretação da modularidade deve ser feita com cuidado, já que seu limite de resolução é determinado pelo tamanho da comunidade [Fortunato & Barthélemy, 2007].

A modularidade é essencialmente a comparação entre o número de arestas de um determinado subgrafo da rede ou comunidade c e o número de arestas no modelo nulo, ou seja, um grafo aleatório de mesmo tamanho e sequência de graus da rede [Fortunato & Barthélemy, 2007]. Assim, um subgrafo é mais próximo de uma boa comunidade quando tem maior modularidade isto é, se o seu número de arestas internas exceder ao número esperado de arestas internas que o mesmo subgrafo teria no modelo nulo. Formalmente, a modularidade pode ser descrita como [Newman & Girvan, 2004]

$$Q = \sum_{c=1}^{m} \left[\frac{l_c}{L} - \left(\frac{d_c}{2L} \right)^2 \right], \tag{2.2}$$

em que l_c é o número de arestas dentro do módulo c, L é o número total de arestas na rede e d_c é o grau total dos vértices no módulo c.

Condutância. A condutância (*conductance*) é outra métrica amplamente utilizada para avaliar a qualidade estrutural de comunidades [Leskovec et al., 2008; Yang & Leskovec, 2015; Zaki & Wagner Meira, 2014]. A condutância mede a qualidade do corte entre um conjunto de vértices e o resto da rede com base no número de arestas fora da comunidade (*inter-cluster conductance*) e no número de arestas dentro da comunidade (*intra-cluster conductance*) [Almeida et al., 2012; Wang & Hopcroft, 2010].

Dado um grafo $G(V, E)$ e um corte s em G, a condutância mede a qualidade de s ou, mais especificamente, quão bem s separa G. Ao fazê-lo, conjuntos de vértices (ou comunidades) com pequenas condutâncias são aqueles que estão densamente conectados internamente e escassamente conectados externamente, sendo, portanto, considerados comunidades de boa qualidade. Com base na condutância, Leskovec et al. [2008] propuseram o método *Network Community Profile (NCP)* que permite obter o melhor *cluster* possível de k vértices e estimar a estrutura da comunidade em grandes redes do mundo real.

2.3.5.2 Avaliação Funcional

Em casos particulares, é possível avaliar a qualidade de comunidades detectadas, comparando-as com o conjunto de metadados de comunidades funcionais ou *ground truth* [Hric et al., 2014; Peel et al., 2017; Zaki & Wagner Meira, 2014]. De acordo com Yang & Leskovec [2015], o *ground truth* é baseado em propriedades particulares do sistema (por exemplo, o departamento de afiliação em uma rede institucional), o que torna possível dividir suas entidades em grupos que compartilham as mesmas propriedades. A representação de tais grupos é feita por um conjunto $P(G)$ de vértices explicitamente rotulados para identificação de suas comunidades.

Para Fortunato [2010], essa forma de verificar a qualidade das comunidades detectadas envolve a definição de um critério para estabelecer quão "semelhante" é a partição fornecida pelo algoritmo em relação à partição que se deseja recuperar (*ground truth*). Os autores apresentam diferentes índices de similaridade, dentre eles *Rand Index* e *Normalized Mutual Information*. Dado um grafo G, um conjunto de comunidades funcionais $P(G)$ e um conjunto de comunidades identificadas $C(G)$, as métricas de similaridade aplicadas às comunidades são capazes de estimar a semelhança entre $C(G)$ e $P(G)$. Assim, selecionamos três índices comumente usados para medir essa semelhança [Fortunato, 2010; Zaki & Wagner Meira, 2014], que são descritos a seguir.

Normalized Mutual Information - NMI. A *NMI* é uma métrica de similaridade da teoria da informação (baseada na dependência mútua entre a entropia associada a uma comunidade identificada e a do *ground truth*). Esta métrica é baseada em uma matriz de confusão, na qual as linhas correspondem ao *ground truth* e as colunas correspondem às comunidades detectadas [Danon et al., 2005], sendo definida como:

$$NMI(X,Y) = \frac{H(X) + H(Y) - H(X,Y)}{(H(X) + H(Y))/2} \tag{2.3}$$

onde H é a função de entropia, X e Y são variáveis aleatórias associadas à comunidade identificada e ao *ground truth*, respectivamente, e $H(X,Y)$ é a entropia conjunta. Seu valor varia de 0 a 1 (quando as comunidades comparadas são idênticas).

Split Join Distance. A *Split Join Distance* mede as sobreposições entre conjuntos de duas partições. Esta métrica é calculada pela soma da distância de projeção entre partições A e B da rede, sendo definida, de acordo com Dongen [2000], como:

$$\rho_A(B) = \sum max|a \cap b| \tag{2.4}$$

onde $|a \cap b|$ indica o número de membros comuns (sobreposição) entre qualquer subconjunto $a \in A$ e $b \in B$ [Zaki & Wagner Meira, 2014].

Rand Index. A métrica denominada *Rand Index* considera a proporção entre o número de concordâncias e o número de discordâncias entre duas partições ou *clusters*. Assim, para medir a semelhança entre dois *clusters*, o número de pares de vértices classificados corretamente (verdadeiros positivos e verdadeiros negativos) em ambos os *clusters* é dividido pelo número total de pares [Rand, 1971]. Essa métrica produz um resultado entre 0 e 1, onde 0 indica que os dois *clusters* não concordam em nenhum par de pontos e 1 indica que eles são exatamente iguais [Zaki & Wagner Meira, 2014].

2.3.5.3 Redução do Viés na Avaliação

Dentre as diferentes estratégias de avaliação do desempenho dos algoritmos de detecção de comunidades, algumas tendem a dar resultados melhores que outras, dependendo da rede [Almeida et al., 2011; Prat-Pérez et al., 2012]. A avaliação de comunidades por meio de métricas que possuem o mesmo princípio utilizado pelo algoritmo escolhido para detectá-las dificulta a explicitação do viés de alguma dessas métricas. Almeida et al. [2011, 2012] exemplificam que algumas métricas de avaliação populares, como a modularidade e condutância, acabam sendo tendenciosas quando aplicadas a grandes comunidades e dão melhores resultados para um número menor de agrupamentos, enquanto outras métricas têm um viés completamente oposto. Eles também afirmam que não existe a "melhor" métrica para avaliação de agrupamentos em grafos. Neste contexto, Yang & Leskovec [2015] complementam que as métricas de qualidade quantificam vários aspectos (em muitos casos mutuamente exclusivos) da estrutura de uma comunidade da rede. Por outro lado, os autores mostram em seus resultados que algumas métricas estruturais estão fortemente correlacionadas. Por isso, a estratégia de avaliação baseada em uma única métrica não é suficiente, o que torna importante a perspectiva de qualidade dada por diferentes métricas para complementar e melhor interpretar e validar as medições.

Neste contexto, além de diversificadas, essas métricas devem constituir pelo menos três estratégias de avaliação baseadas em pressupostos distintos sobre a qualidade de uma comunidade, de modo que potenciais vieses possam estar em direções opostas [Leão et al., 2018]. Assim, em caso de resultados que indiquem decisões divergentes sobre qual é o melhor conjunto de comunidades detectadas, torna-se possível obter um consenso a partir da decisão que prevaleça entre as estratégias.

Por exemplo, apesar do uso combinado de métricas como modularidade e condutância permitir a obtenção de múltiplas evidências sobre a qualidade de uma comunidade [Yang & Leskovec, 2015], o consenso obtido pode ser tendencioso no sentido da qualidade estrutural [Zaki & Wagner Meira, 2014]. Isto significa que, mesmo sendo consideradas boas métricas, elas avaliam a qualidade das comunidades apenas com base no aspecto topológico e, portanto, com pressuposto único sobre a qualidade de uma comunidade. Idealmente, pode ser feita, além da avaliação estrutural, a comparação com outro referencial de qualidade como uma avaliação funcional ou pela comparação com um *baseline*.

Naturalmente, pode existir alguma divergência entre evidências de fontes distintas, o que levaria a uma estimativa de consenso menor do que a que seria obtida pelo uso de métricas baseadas em um único pressuposto de qualidade. Nesse sentido, consi-

deramos também que o consenso obtido a partir de pressupostos distintos ou até mesmo divergentes é mais representativo da qualidade real da estrutura de uma comunidade. Por exemplo, Hric et al. [2014] mostram em seus resultados que existe um limiar de separação entre comunidades estruturais e o *ground truth* e, por isso, concluem que a modelagem atual da estrutura de comunidades[4] deve ser substancialmente modificada ou que as comunidades funcionais não são recuperáveis apenas pela topologia. Utilizar múltiplas estratégias distintas para avaliar a qualidade de comunidades permite uma interpretação transversal e um melhor mapeamento da abrangência e profundidade dos resultados obtidos. Além disso, possibilita que algum viés nas métricas ou nos dados possa ser estimado pela análise de diferentes fontes de evidência.

[4]Entende-se que a estrutura de comunidade é modelada a partir de propriedades topológicas.

Capítulo 3

Arcabouço Proposto

Uma comunidade é uma das estruturas mais representativas de uma rede social e é constituída por muitos laços ou relacionamentos sociais fortes [David et al., 2010; Fortunato, 2010]. Por isso, é esperado que uma comunidade possa ser melhor extraída de uma rede quando a sua estrutura for representada apenas por relacionamentos sociais e livres de ruído [Abufouda & Zweig, 2017]. Neste capítulo, apresentamos nosso arcabouço proposto para filtrar relacionamentos e melhorar os resultados da tarefa de detecção de comunidades. Também são descritas as técnicas de detecção de comunidades utilizadas para a etapa de avaliação.

A principal ideia por trás do arcabouço proposto é remover das redes o conjunto de interações que correspondem a uma ou mais classes de relacionamentos. A Figura 3.1 detalha as principais etapas que resumem o funcionamento do nosso arcabouço para remoção de relacionamentos na classe aleatória, que são: (i) classificação dos relacionamentos a partir do fluxo de interações, (ii) remoção de relacionamentos aleatórios e (iii) construção da rede estática S que será usada como entrada para as técnicas de detecção de comunidades.

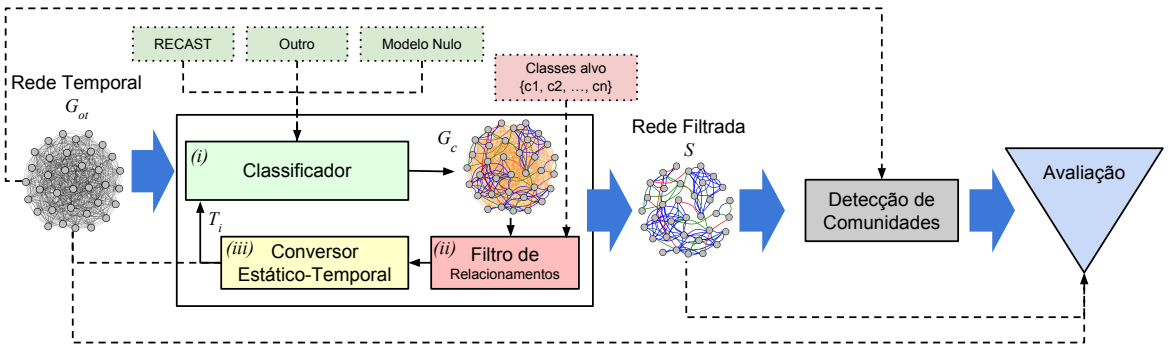

Figura 3.1: Visão geral do arcabouço que obtém uma rede filtrada estática S.

Require: G_{ot}: rede temporal (original)
1: c {classes selecionadas para filtrar}
2: $i = 1$, K {limite de iterações}
3: $T_i \leftarrow G_{ot}$
4: **while** $i < K$ **do**
5: $G_c \leftarrow classificar(T_i)$
6: $S, R \leftarrow filtrar(G_c, c)$
7: **if** $converge(G_c, S)$ **then**
8: **break**
9: **end if**
10: $T_{i+1} \leftarrow converter(S, T_i)$
11: $i = i + 1$
12: **end while**
13: **return** S

Algorithm 1: Filtragem de Relacionamentos.

É importante notar que, na etapa de classificação dos relacionamentos, resultados diferentes podem ser produzidos para um dado relacionamento quando a sequência filtrada de interações é usada como entrada pela segunda vez. Por isso, após o passo (*iii*), o passo (*i*) é realizado novamente usando a rede filtrada como entrada. Este ciclo é interrompido quando o passo (*ii*) não remove mais nenhum relacionamento. Então, quando não há mais relacionamentos aleatórios, obtemos uma rede estática S que é composta apenas de relacionamentos sociais. Esse processo permite que apenas os relacionamentos de interesse (i.e., aqueles que são sociais) estejam disponíveis como entrada para alguma técnica de detecção de comunidades. Após a filtragem de relacionamentos, avaliamos os resultados através das estratégias de avaliação funcional e estrutural, além da comparação com um *baseline*, conforme descrito nas seções seguintes.

O Algoritmo 1 apresenta as principais etapas de nosso processo de filtragem de relacionamentos sociais, que recebe como entrada uma sequência de interações e os seguintes parâmetros de configuração: o classificador de relacionamentos e o conjunto de classes que devem ser removidas. Em cada iteração do algoritmo, os relacionamentos são classificados (linha 5), resultando em uma rede de relacionamentos rotulados $G_c = (V_c, R_c)$. Em seguida, a função *filtrar* remove de G_c os relacionamentos aleatórios (linha 6). Se na verificação de convergência (linha 7) ainda existirem arestas rotuladas como aleatórias, é executada novamente a função que constrói a rede temporal T_{i+1} (linha 10) que servirá de entrada para uma nova iteração. Observe que a função *filtrar* constrói dois subgrafos induzidos pelas arestas de G_c: o grafo filtrado $S{=}(V_S, E_S)$ e o grafo residual $R{=}(V_R, E_R)$. Cada subgrafo induzido por arestas é um subconjunto das arestas do grafo G_c que contém vértices de V_S em suas extremidades. E_S é o conjunto de arestas rotuladas como *sociais* em E_c. Por sua vez, E_R é o complemento de E_S, ou seja, o conjunto de arestas rotuladas como *aleatórias*.

3.1 Ruídos

Em redes sociais é natural que existam relacionamentos aleatórios que causam ruído em sua estrutura [Abufouda & Zweig, 2017; Vaz de Melo et al., 2015]. Isto ocorre porque os pares de vértices envolvidos nesses relacionamentos possuem pouca probabilidade de interagirem novamente[1] [David et al., 2010; Kossinets & Watts, 2006]. Assim, definimos ruído como a pertubação estrutural causada pela presença de interações aleatórias que obscurecem ou não são especificamente significativas para a estrutura de uma comunidade. Por isso, quantificamos ruído em função do número de relacionamentos aleatórios. Note que consideramos apenas a relação de causalidade e, portanto, a probabilidade de um relacionamento aleatório causar ruído não está associada à semântica desse relacionamento em um domínio específico de uma rede ou a outra aplicação além da detecção de comunidades. Isto significa que uma aresta que causa ruído estrutural em uma comunidade não necessariamente causará algum outro tipo de ruído, por exemplo, para a disseminação de informação na rede, onde relacionamentos aleatórios geralmente têm um efeito oposto ao tipo de ruído que definimos e utilizamos nesta dissertação [Leão et al., 2017b].

Ao configurarmos o arcabouço de filtragem proposto para remover os relacionamentos aleatórios, conseguimos obter uma rede livre de ruído. Especificamente, consideramos um cenário em que a representação eficaz para uma rede social é uma rede de agregação temporal G_{ot}, em que as sequências de interações são agregadas em períodos discretos de tempo. Então, como é usual, construímos uma rede de relacionamentos G_{os} pela agregação das interações na rede G_{ot} ao longo de todo o tempo e usamos essa rede como entrada para alguma técnica de detecção de comunidades.

Nossa primeira hipótese é que, se pudermos identificar na rede G_{ot} pares de vértices v_i e v_j correspondentes a entidades que interagiram por acaso (ou que seja pouco provável que venham a interagir novamente), podemos remover todas as interações entre v_i e v_j antes de construir a rede S, melhorando assim a qualidade da representação estática da rede G_{ot} e, consequentemente, permitindo maior qualidade na detecção das comunidades dessa rede. Neste contexto, o objetivo principal do uso do arcabouço proposto é reduzir o erro ao associar vértices a comunidades.

Durante o processo de filtragem é possível que algum vértice possua todos os seus relacionamentos classificados como aleatórios. Assim, na etapa de filtragem, consequentemente vértices assim são desconectados da rede. Esses vértices são considerados

[1] Dentre os aspectos que podem ser considerados para medir a probabilidade de pares de indivíduos interagirem novamente, Kossinets & Watts [2006] e David et al. [2010] destacam o número de conhecidos mútuos e o número de interações entre eles.

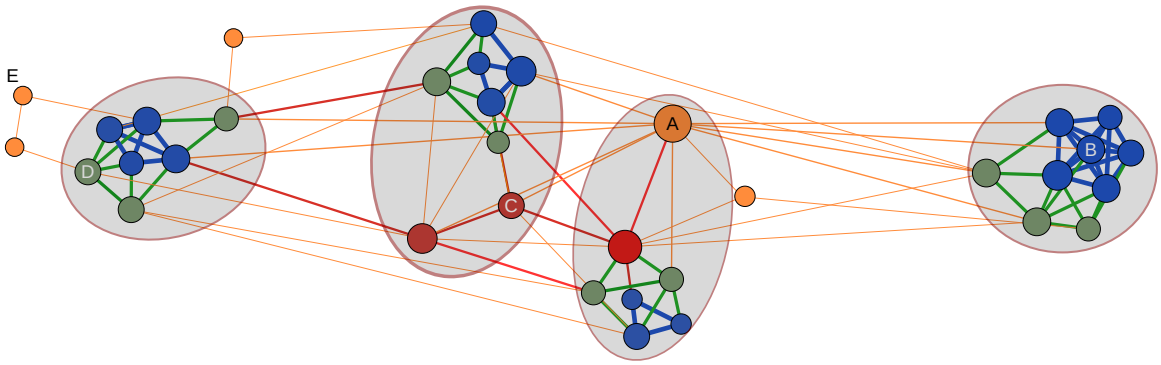

Figura 3.2: Exemplo de uma rede de relacionamentos. Arestas de mesma cor identificam uma classe de relacionamentos. Vértices (entidades) de mesma cor representam a sua classe de relacionamento predominante [Leão et al., 2017b], envolvidos por círculos que indicam as comunidades a que pertencem.

"violadores" da estrutura da rede, pois não se relacionam significativa e distintamente em uma ou mais comunidades. Na Figura 3.2 é exemplificada uma rede com relacionamentos de diferentes classes, compondo comunidades. Além disso, é possível verificar a presença de vértices "violadores" fora das comunidades, como o vértice *E*.

3.2 Detecção de Comunidades

Com base no estado da arte em detecção de comunidades, selecionamos os algoritmos listados na Tabela 3.1 para avaliar a sua eficácia antes e depois da remoção de ruído. Assim, buscamos responder à nossa primeira questão de pesquisa (Q1) sobre o efeito da filtragem de relacionamentos aleatórios para algoritmos de detecção de comunidades. Esses algoritmos são descritos sucintamente a seguir[2].

Edge Betweenness (Girvan–Newman). Este algoritmo descobre comunidades em redes dividindo vértices em subgrupos densamente conectados [Newman & Girvan, 2004]. Baseia-se na remoção iterativa de arestas da rede de acordo com seu valor de intermediação, que é recalculado após cada remoção. Em suma, a intermediação é uma medida de centralidade de uma aresta em uma rede baseada no número de caminhos mais curtos que passam por essa aresta. A estratégia desse algoritmo resume-se a remover primeiro as arestas mais centrais, que são aquelas que normalmente conectam as maiores comunidades.

Greedy Optimization of Modularity. É um algoritmo baseado na maximização da modularidade usando uma abordagem gulosa [Clauset et al., 2004]. Em um primeiro

[2]Alguns algoritmos não executaram em um tempo razoável sobre os conjuntos de dados maiores usados em nossos experimentos. Por isso, esses algoritmos não aparecem em todas as nossas análises.

Tabela 3.1: Algoritmos para detecção de comunidades.

Abordagem	Algoritmo	ξ	O	Referência
Maximização da modularidade	Louvain Modularity (LM)	D	$VlogV$	[Blondel et al., 2008]
	Greedy Optimization of Modularity (GOM)	D	$Vlog^2V$	[Clauset et al., 2004]
	Leading Eigenvector (LE)	D	V^2logV	[Newman, 2006a,b]
Custo de trajetória	Infomap (IM)	N	$VlogV$	[Rosvall et al. 2011]
Processo dinâmico	Label Propagation (LP)	N	V	[Raghavan et al., 2007]
Remoção de arestas entre comunidades	Edge Betweenness (EB)	D	V^3	[Newman & Girvan, 2004]
Similaridade de vértices	Walktrap (WT)	N	V^2logV	[Pons & Latapy, 2005]

ξ: modelo de estado do algoritmo (D-determinístico ou N-não determinístico); O: ordem de complexidade de tempo (limite assintótico superior) calculada sob o pressuposto de que o grafo é esparso.

passo, ele identifica uma estrutura hierárquica de comunidades. Então, é feita uma partição na hierarquia de forma a maximizar globalmente a modularidade.

Infomap. Através deste algoritmo, as comunidades são descobertas aplicando a técnica de passeio aleatório para mapear o fluxo de informações através de uma rede. Infomap agrega em uma comunidade um grupo de vértices através dos quais a informação flui rápida e facilmente entre eles [Rosvall & Bergstrom, 2011]. Para isso, codifica a descrição de uma trajetória de passeio aleatório. Então, encontra comunidades quando minimiza o comprimento estimado dessa descrição.

Label Propagation. É um método estocástico de detecção de comunidades com base na propagação de rótulos entre vértices [Raghavan et al., 2007]. Cada vértice é inicializado aleatoriamente com um rótulo. Então, os vértices têm seus rótulos substituídos iterativamente por aquele da maioria dos seus vizinhos. Desta forma, grupos densamente conectados formam um consenso em seus rótulos indicando que eles participam da mesma comunidade.

Leading Eigenvector. Este método separa os vértices em comunidades considerando o autovetor da matriz de modularidade do grafo [Newman, 2006a,b]. A matriz de

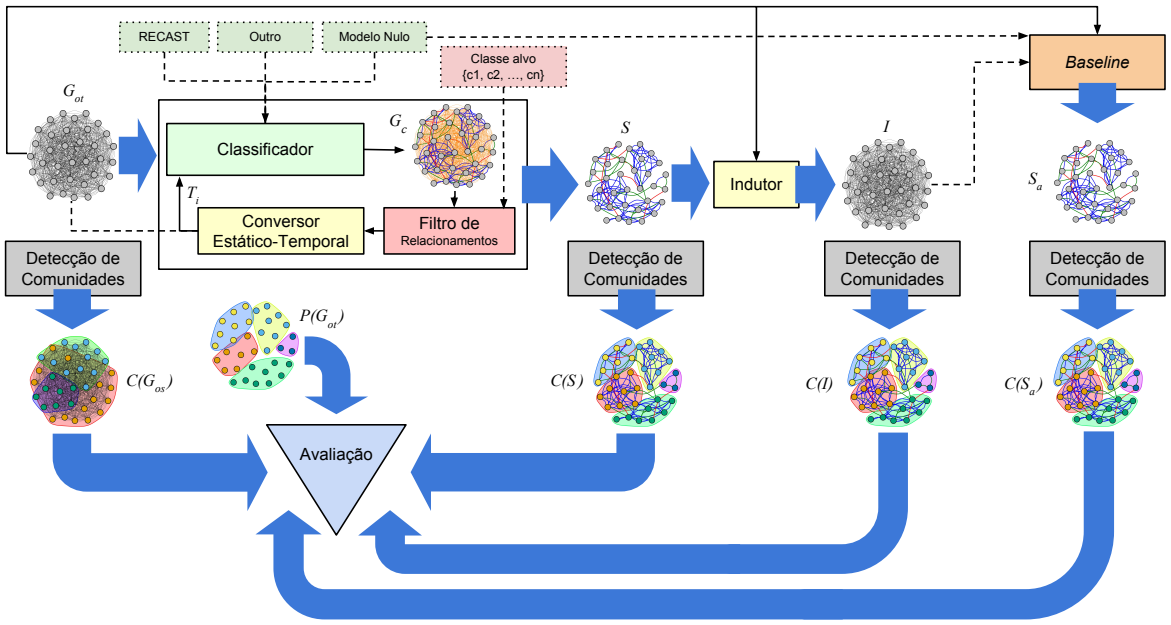

Figura 3.3: Detalhamento da etapa de coleta de evidências sobre a eficácia da filtragem de relacionamentos sociais para a melhoria da detecção de comunidades.

modularidade desempenha um papel na detecção de comunidade semelhante ao de uma matriz Laplaciana na partição de um grafo. Assim, diferentes sinais dos elementos no autovetor determinam comunidades distintas ou, em outros casos, que não há estrutura de comunidade subjacente.

Louvain. Este é um método guloso para otimização da modularidade que constrói uma estrutura hierárquica das comunidades em duas etapas. Primeiro, agrupa os vértices vizinhos em comunidades "pequenas" a fim de otimizar a modularidade localmente. Em seguida, ele constrói a estrutura hierárquica pela agregação de cada vértice de um grupo em um vértice de um novo grafo [Abrahao et al., 2012; Blondel et al., 2008].

Walktrap. Este é um algoritmo aglomerativo que calcula a estrutura de comunidade de uma rede com base em uma métrica de similaridade entre vértices [Pons & Latapy, 2005]. Essa métrica também é baseada no passeio aleatório, que capta naturalmente a estrutura de comunidade em uma rede e pode ser eficientemente computada. Assim, dois vértices são mais similares quanto menor for o caminho entre eles.

3.3 Estratégias de Avaliação

Conforme a Figura 3.3, após a etapa de filtragem de relacionamentos, avaliamos a qualidade da rede resultante para a tarefa de detecção de comunidades. Considerando o cenário diversificado e as divergências nas abordagens de detecção e nas definições de

comunidade, buscamos responder à nossa segunda questão de pesquisa (Q2) através de uma avaliação de qualidade das comunidades geradas quando a detecção é feita a partir de uma rede filtrada. Essa avaliação vai além da medição da qualidade das comunidades obtidas por cada algoritmo, mas também compara as melhorias alcançadas por algoritmos com abordagens distintas. Mais importante, também analisamos a qualidade das comunidades na rede original (intacta e com ruído) e a comparamos com a qualidade das comunidades obtidas da rede filtrada.

A coleta de evidências sobre a eficácia do nosso arcabouço se inicia antes do processo de filtragem e vai até o seu final, conforme resumimos a seguir. Inicialmente, as sequências de interações são agregadas para construir a rede de relacionamentos G_{os}. Em seguida, executamos cada um dos algoritmos listados na Tabela 3.1 sobre esta rede e medimos a qualidade das comunidades detectadas $C(G_{os})$ usando as métricas de avaliação estrutural (dentre elas, modularidade e condutância). Então, medimos a similaridade entre essas comunidades detectadas através das métricas selecionadas para este propósito. Também é feita a comparação entre a similaridade das comunidades em $C(G_{os})$ e as comunidades funcionais $P(G_{os})$, quando o *ground truth* está disponível.

Depois disso, usamos o nosso arcabouço para filtrar a rede G_{ot}. Então, obtemos a sequência de interações filtradas T_i que, em seguida, são agregadas para construir a rede de relacionamentos filtrada S. Nesta etapa também é gerado um subgrafo $I = G_{ot}[V_S]$, induzido[3] da rede G_{ot} por vértices a partir de S. Ao final dessa etapa, todos os algoritmos também são executados sobre S e as comunidades obtidas $C(S)$ também têm suas estruturas avaliadas e comparadas entre si e com o *ground truth*.

Na etapa final da avaliação, comparamos as características estruturais e funcionais da rede original (G_{os}) com as características da rede filtrada (S) e registramos o ganho obtido em cada característica. Paralelamente, todo o processo executado até essa etapa é repetido com o modelo nulo e com o *baseline* e os seus resultados são utilizados na análise de evidências.

Assim, em nossa avaliação, verificamos uma melhoria na detecção de comunidades através do consenso entre as evidências coletadas durante o processo de filtragem. Especificamente, foram selecionados métodos de avaliação compostos pelas métricas apresentadas na Seção 2.3.5. Tais métodos foram agrupados de forma a compor três estratégias de avaliação com pressupostos distintos sobre a qualidade de uma comunidade, como descrito a seguir.

[3]O grafo I representa a rede filtrada por vértices aleatórios, ou seja, a rede construída pela remoção dos vértices que violam a estrutura de comunidade e que foi utilizada para evidenciar se, para a detecção de comunidades, a remoção de arestas aleatórias é melhor do que a remoção de vértices aleatórios.

Avaliação da qualidade estrutural. A primeira estratégia considera a qualidade de uma comunidade determinada por suas características estruturais. Métricas que medem a qualidade estrutural de uma comunidade, como a condutância [Clauset, 2005] e a modularidade [Newman & Girvan, 2004], permitem quantificar o quanto um agrupamento da rede se parece com estruturas de comunidades [Yang & Leskovec, 2015]. Utilizando essas métricas, a avaliação foi conduzida para as tuplas $<$*método de filtragem*, *configuração do método*, $C_{A_i}(X)>$, onde $C_{A_i}(X)$ é o conjunto de comunidades detectadas por cada algoritmo A_i sobre cada uma das redes em $X = \{G_{os}, S, I, S_a\}$. Para as tuplas que envolvem eventos estocásticos no método ou no algoritmo, os experimentos foram executados com pelo menos 30 replicações para estimar valores médios. Quando a rede é de natureza sintética, as replicações também foram feitas sobre 20 instâncias geradas a partir dos mesmos parâmetros do modelo de rede simulada. Além das métricas de qualidade, utilizamos da visualização, da contagem do número de comunidades detectadas, da variância desse número e de métricas de caracterização de redes sociais para analisar as alterações estruturais gerais das redes.

Também obtivemos evidências sobre a melhoria na qualidade das comunidades através do aumento no consenso entre os algoritmos sobre as comunidades que devem ser detectadas. Essas evidências permitem estimar o quanto o ruído interfere na detecção das mesmas comunidades por algoritmos distintos. Para isso, medimos a similaridade entre as comunidades obtidas por pares de técnicas distintas A_i e A_j na rede original G_{os} e o ganho $g^X_{A_i,A_j}$ obtido sobre essa similaridade após a filtragem da rede. Esse ganho foi calculado pela diferença entre a similaridade calculada sobre a rede original G_{os} e a similaridade calculada em cada rede filtrada em X, ou seja, $g^X_{A_i,A_j} = sim(C_{A_i,A_j}(X)) - sim(C_{A_i,A_j}(G_{os}))$. Então, construímos matrizes de similaridade em que cada elemento contém o valor de similaridade na rede original e o ganho para um dos possíveis pares A_i e A_j, como no exemplo da Figura 3.4. Assim, a construção dessas matrizes foi repetida para cada uma das métricas de similaridade.

Na Figura 3.4 são exemplificados os valores de similaridade de cada tupla (A_1, A_2) na rede original (Figura 3.4a) e na rede filtrada S (Figura 3.4b) e o ganho (Figura 3.4c), obtido da comparação entre os algoritmos *Label Propagation* (LP), *Walktrap* (WT) e *Infomap* (IM) sobre a rede *High School*. Com base nessas matrizes, também é possível obter o ganho em similaridade entre comunidades distintas, detectadas por um mesmo algoritmo não determinístico, como LP, IM ou WT, o que indica que a remoção do ruído aumenta a precisão desses algoritmos. Ademais, é importante notar na matriz resultante o ganho em similaridade entre as comunidades detectadas por algoritmos distintos. Nesse caso, o seu valor de ganho indica o quanto aumentou o consenso entre diferentes algoritmos em relação a quais comunidades devem ser detectadas.

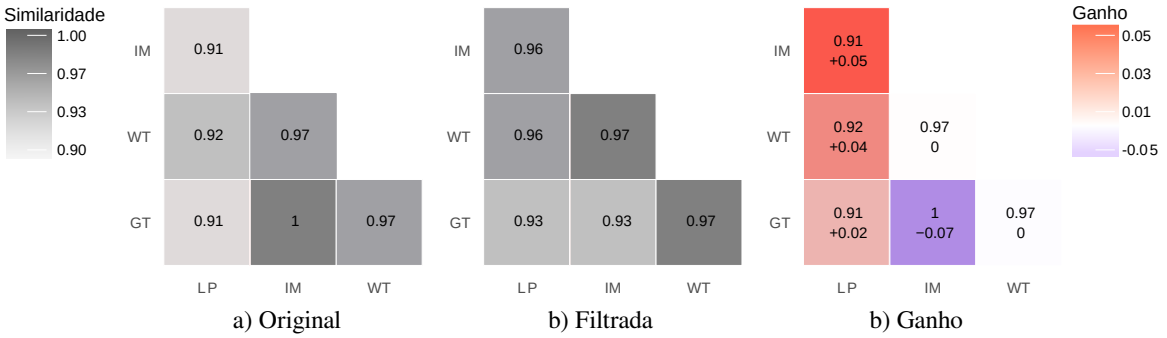

Figura 3.4: Exemplo de construção da matriz de consenso (b) para a rede *High School* através da métrica *NMI*. Em cada célula temos: a) similaridade na rede original; b) similaridade na rede filtrada; c) similaridade na rede original e ganho na rede filtrada.

Avaliação da qualidade funcional. Nesta estratégia, a qualidade de uma comunidade detectada é medida pela sua similaridade com o *ground truth*. As comunidades que compõem um *ground truth* foram explicitamente rotuladas com um identificador. Assim, os vértices com atributos de valores iguais são associados ao mesmo rótulo (comunidade). Em seguida, fazemos a comparação entre as comunidades detectadas e o conjunto de comunidades rotuladas da respectiva rede antes e após o processo de filtragem. A Figura 3.4 exemplifica as matrizes de similaridade entre os algoritmos *Label Propagation* (LP), *Walktrap* (WT) e *Infomap* (IM) em relação ao *ground truth* (GT). Assim, a Figura 3.4a apresenta a matriz de similaridade entre o *ground truth* e a rede original, enquanto que Figura 3.4b apresenta a matriz de similaridade entre o *ground truth* e a rede filtrada usando uma das métricas de similaridade já mencionadas. Finalmente, a Figura 3.4c apresenta o ganho de similaridade entre as duas medidas. Essas redes rotuladas foram obtidas a partir de duas fontes distintas: o *ground truth* de redes reais, quando disponível e o *ground truth* fornecido como entrada para um gerador de modelos sintéticos de rede sobre os quais temos controle prévio de quais grupos representam as comunidades.

Avaliação da qualidade relativa a um *baseline*. A qualidade do método e do modelo utilizados em nosso arcabouço pode ser avaliada pela comparação com outros métodos e modelos propostos para a mesma finalidade. Em nosso arcabouço, o principal modelo de força dos laços utilizado considera a regularidade das interações [Vaz de Melo et al., 2015]. Para verificar que esse modelo permite identificar ruído com chance maior que a aleatória, comparamos com um modelo nulo, baseado em um método estocástico de filtragem. Além disso, foi utilizado como *baseline* o método de filtragem de ruído baseado na remoção de vértices violadores proposto por Wen et al. [2011].

Tabela 3.2: Configurações do Gerador de Mobilidade

Redes de Relacionamento usadas como Entrada	
CSC	Comunidades sintéticas (cliques)
CSS	Comunidades sintéticas com sobreposições
CRR	Comunidades com relacionamentos reais

Pelo cruzamento dos resultados obtidos pelas estratégias de avaliação estrutural, funcional e de comparação com um *baseline*, obtivemos linhas distintas de evidência sobre a melhoria da qualidade das comunidades. Isto permite avaliar de forma robusta o nosso arcabouço. Além disso, mesmo considerando múltiplas definições para a estrutura de comunidade, essa avaliação permite obter-se uma certeza consensual sobre a eficácia na filtragem da rede. Note que essa combinação de estratégias de avaliação sobre redes de domínios distintos nos permitiu analisar a variação do ganho obtido pela filtragem de relacionamentos aleatórios entre os algoritmos de detecção utilizados e sobre tipos de interação diferentes para responder à nossa terceira questão de pesquisa (Q3).

Avaliação em redes simuladas. Uma rede sintética pode ser obtida a partir de um modelo de rede social. Tais modelos devem refletir de forma realista as propriedades de redes sociais [Treurniet, 2014]. Por exemplo, em cenários de mobilidade, é possível construir uma rede com interações sintéticas com características como duração e tempo entre as interações, estrutura de grupos (comunidades), regularidade espacial, dentre outras. A partir do gerador GRM [Nunes et al., 2017], obtivemos sequências de interações fornecendo como entrada uma rede estática de relacionamentos em que sabemos previamente quais são as comunidades funcionais.

Utilizamos três cenários para gerar a rede de interações sintéticas (Tabela 3.2). Primeiro, fornecemos relacionamentos sintéticos em que os grupos foram construídos entre pessoas que pertencem à mesma comunidade. As comunidades são parametrizadas em número e tamanho com base em *ground truths* reais, formando relacionamentos completos entre membros da mesma comunidade (cliques). Fornecemos esses cliques como entrada para o gerador. No segundo experimento, repetimos o que foi feito no primeiro e perturbamos essas redes inserindo relacionamentos aleatórios, obtendo-se sobreposições entre as comunidades. No terceiro experimento, fornecemos redes de relacionamentos reais para o gerador em que também se conhece o *ground truth* das comunidades. Então, para cada rede G_{ot} obtida em cada cenário, avaliamos a melhoria na detecção de comunidades obtida pelo arcabouço de filtragem de relacionamentos. Isso permitiu avaliar a qualidade obtida ao detectar as comunidades e verificar, quantitativamente, o quanto as relações aleatórias prejudicam essa detecção.

Capítulo 4

Resultados Experimentais

Neste capítulo, apresentamos os resultados dos experimentos realizados com o arcabouço proposto, como também as redes e configurações utilizadas nos experimentos. Também são demonstradas as alterações nas características estruturais entre as redes sociais que permitiram estimar o efeito do ruído na estrutura das comunidades identificadas. Além disso, detalhamos os resultados das diferentes estratégias de avaliação que evidenciaram que a aplicação do nosso arcabouço melhora a qualidade das comunidades detectadas.

Entretanto, é importante ressaltar que durante os experimentos iniciais verificou-se um grande consumo de recursos computacionais por alguns dos algoritmos adotados, o que inviabilizou a sua utilização nas redes utilizadas (Tabela 4.1) que, de acordo com o referencial de Kumpula & Kaski [2008] e Pollner et al. [2012], são consideradas como de grande porte.

Assim, todos os experimentos foram repetidos devido à variabilidade associada ao tempo de execução e às estimativas de valores obtidos por algoritmos não determinísticos. Em particular, para os algoritmos *Walktrap* e *Edge-Betweenness* as repetições não puderam ser registradas ou não terminaram dentro do limite de tempo de 45 dias. Nesses casos, resultados incompletos não foram considerados na análise.

4.1 Caracterização das Redes Utilizadas

Redes sociais podem ser diferenciadas pela natureza de seus relacionamentos. Assim, a diversidade de tipos de relacionamento é observada mesmo considerando um único domínio e o envolvimento de um mesmo tipo de entidade. Por exemplo, em uma rede social, relacionamentos podem se referir a amizades entre colegas de classe ou envolver profissionais que apenas pertencem ao mesmo departamento de uma instituição ou

Tabela 4.1: Caracterização das redes sociais.

Domínio	*Rede*	Período	$\|V\|$	$\|E\|$	Δ	D	CC
Colaboração Científica	*APS*	13 anos	181k	852k	305	0.5	0.33
	PubMed	16 anos	444k	5.5M	4869	0.6	0.36
	DBLP	15 anos	945k	3.8M	1413	0.1	0.16
	arXiv	25 anos	33k	180k	424	3.3	-
Mobilidade	*Dartmouth*	8 semanas	1.1k	25k	236	410	0.51
	USC	8 semanas	2.5k	160k	652	510	0.49
Propagação de doença	*High School*	5 dias	327	5818	87	1.1k	0.44
Amizade	*Primary School*	32 horas	242	8317	134	2.8k	0.48
Comunicação	*Enron*	4 anos	87k	321k	1566	0.8	0.07
	Email-Eu-core	803 dias	986	25k	211	513	0.27
Simuladas	*CSC, CSS e CRR*	-	$\approx$1k	$\approx$13k	$\approx$78	$\approx$267	$\approx$0.29

$|V|$: número de vértices; $|E|$: número de arestas; Δ: grau máximo; D: densidade (x10^{-4}); CC: coeficiente de agrupamento. O grau mínimo em todas as redes é 1. Estas e outras propriedades dessas redes estão disponíveis com maiores detalhes em: `http://cnet.jcloud.net.br/`.

fazem parte de um mesmo grupo de trabalho [Barrat et al., 2008]. Nesta seção apresentamos as redes sociais que utilizamos nos experimentos realizados nesta dissertação. Além disso, caracterizamos a estrutura de comunidade identificada em cada uma dessas redes e exemplificamos para uma delas.

Redes Sociais Reais. Inicialmente, modelamos como redes de agregação temporal as redes sociais de colaboração científica[1] [Brandão & Moro, 2017], as redes de mobilidade de campus universitário[2] [Vaz de Melo et al., 2015], redes de *e-mails* derivadas da comunicação entre colaboradores da Enron e de uma instituição de pesquisa europeia[3] [Leskovec et al., 2007; Rossetti & Cazabet, 2017] e redes de contatos entre membros de escolas primárias e secundárias[4] [Gemmetto et al., 2014]. A Tabela 4.1 apresenta uma caracterização geral dessas redes. Nas redes de colaboração científica, os vértices representam pesquisadores e há uma aresta ligando dois pesquisadores se eles são coautores de um mesmo artigo. Nas redes de mobilidade, os vértices representam usuários

[1]Conjuntos de dados obtidos de `http://homepages.dcc.ufmg.br/~mirella/projs/apoena/`: APS: rede de coautoria de membros da *American Physical Society*; PubMed: rede de coautoria de artigos disponíveis na MEDLINE; DBLP: rede de coautoria de artigos apresentados em conferências de ciência da computação disponíveis na DBLP; arXiv: rede de coautoria de artigos obtida de `https://www.kaggle.com/neelshah18/arxivdataset/`

[2]*Dartmouth College* e USC, obtidos de `https://crawdad.org/`.

[3]Conjuntos de dados de *e-mails* obtidos de `https://snap.stanford.edu/data/`.

[4]Conjuntos de dados obtidos de `http://www.sociopatterns.org/datasets/`.

de um campus universitário (por exemplo, estudantes ou membros do corpo docente de uma universidade) e há uma aresta entre dois indivíduos se ambos estiverem conectados a um determinado ponto de acesso *Wi-Fi* ao mesmo tempo. Finalmente, os vértices das redes de *e-mails* são membros de uma instituição e há uma aresta entre eles caso tenham trocado *e-mails*.

Redes Sintéticas. Como mencionado na Seção 3.3, utilizamos também redes simuladas para complementar a nossa estratégia de avaliação. O GRM [Nunes et al., 2017] apresentou-se como o modelo mais completo para representar a mobilidade com características de grupos (comunidades). Contudo, não são comuns estudos que apresentam a comparação direta entre as estruturas de comunidades geradas por esse modelo e as estruturas de comunidades funcionais ou estruturais detectadas por algoritmos do estado da arte. Assim, além de fazer essa comparação para avaliar o nosso arcabouço, demonstramos também, pela comparação com outros métodos convencionais de avaliação da qualidade de comunidades, que um modelo representa satisfatoriamente a estrutura de uma comunidade estática.

Metadados dos *Ground Truths*. Nas redes de colaboração científica, as comunidades são identificadas pelos veículos em que os pesquisadores publicam predominantemente. Assim, extraímos os identificadores dos periódicos nos quais os pesquisadores publicaram os seus artigos e que caracterizam as redes consideradas. Nas redes simuladas, os *ground truths* se baseiam nas três configurações de modelos apresentados na Tabela 3.2: comunidades sintéticas (CSC), comunidades com sobreposições sintéticas (CSS) e comunidades com relacionamentos reais (CRR).

Para exemplificar a construção e caracterização realizada usando um *ground truth*, utilizamos a rede de colaboração APS. O *ground truth* para tal rede é sumarizado na Tabela 4.2 e foi obtido a partir do identificador do periódico em que cada membro da rede publica predominantemente. É importante observar na Figura 4.1b que, em duas porções distintas ampliadas da rede da Figura 4.1a, os vértices de diferentes comunidades funcionais não estão na mesma comunidade estrutural identificada pelas cores da Figura 4.1c. Essa característica se deve à diversidade de áreas nas quais um mesmo pesquisador da Física publica (sobreposições entre áreas). Ademais, isso leva a uma baixa semelhança entre a topologia estática da rede e a área em que cada pesquisador publica predominantemente. A aparente dissimilaridade é confirmada pelos valores baixos obtidos pela maioria das medidas de similaridade utilizadas na comparação entre as comunidades detectadas e o *ground truth*.

Tabela 4.2: Comunidades reais da rede APS.

N	Título do Periódico (comunidade)	*Disjuntas*	*Sobreposições*
1	Review A	26K	86K
2	Review B	67K	195K
3	Review C	8K	24K
4	Review D	21K	77K
5	Review E	38K	96K
6	Physical Review Letters	31K	90K
7	Accelerators and Beams	4K	9K
8	Physics Education Research	0.3K	688
9	Physical Review X	331	860
10	Reviews of Modern Physics	448	1459

Participações disjuntas: Número de participantes apenas na área em que participam mais frequentemente (comunidade disjunta);
Sobreposições: Número de participantes na comunidade representada por várias áreas por membro (sobreposição de comunidades);
Comunidade derivada (funcional): Physical Review A - física atômica, molecular, óptica e quântica; Physical Review B - matéria condensada e física dos materiais; Physical Review C - física nuclear; Physical Review D - partículas, campos, gravitação e cosmologia; Physical Review E - estatística, não linear, biológica e matéria mole; e Physical Review X - física interdisciplinar.

Tabela 4.3: Medidas de modularidade da rede APS.

Algoritmo	Ground Truth	LM	GOM	LE	IM	LP	EB
Repetições	1	6	5	1	50	156	1
Modularidade	0.53	0.81	0.70	0.35	0.13	0.66	0.80
Variância	0	0	0	0	2.10^{-8}	8.10^{-6}	0
Nº de Comunidades	10	5,085	6,595	4,963	73,625	16,806	-
Variância	0	0	0	0	3,000	7,838	-

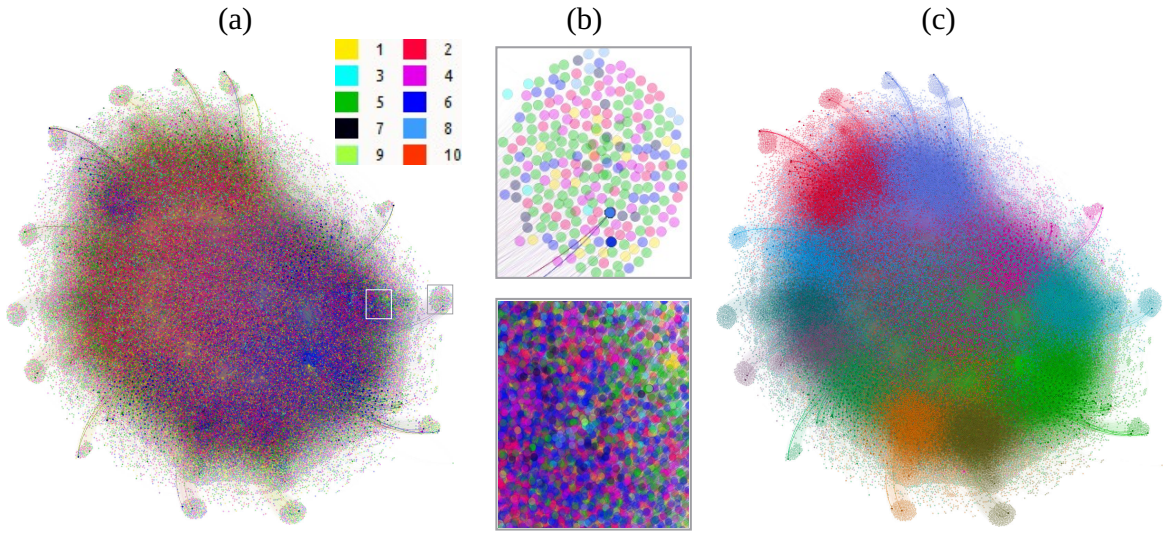

Figura 4.1: Estrutura da rede APS: (a) comunidades reais (as 10 maiores comunidades correspondem a 100% dos vértices); (b) subgrafo coeso com vértices na cor que identifica sua comunidade funcional; (c) comunidades detectadas pelo algoritmo *Louvain* [Blondel et al., 2008] (as 10 maiores comunidades correspondem a 98 % dos vértices).

A Figura 4.2 apresenta três medidas de similaridade entre as comunidades da rede APS. Na coluna GT são apresentados os valores de similaridade entre as comunidades funcionais (*ground truth*) e as estruturais. Nas demais colunas são apresentados os valores de similaridade entre as comunidades estruturais detectadas por algoritmos distintos. Note que, diferentemente das outras métricas que indicam a similaridade através de valores em um intervalo entre 0 e 1, a métrica *Split Join Distance* possui valores absolutos e que indicam maior similaridade quando seu valor é menor. Assim, na Figura 4.2 é evidenciado que diferentes algoritmos detectam comunidades mais parecidas entre si do que com as comunidades funcionais (GT), o que confirma que originalmente o *ground truth* da rede APS não é compatível com a sua estrutura. Os valores de modularidade na Tabela 4.3 também evidenciam essa incompatibilidade entre módulos extraídos e comunidades funcionais, mostrada na Figura 4.1c.

4.2 Classificação

O problema de classificação de relacionamentos em redes sociais consiste em atribuir um rótulo do conjunto L a cada par de vértices (v_i, v_j) que possui alguma interação em G_{ot}. Como antecipado, nesta dissertação consideramos aspectos topológicos e temporais para medir a força dos relacionamentos e então determinar qual rótulo será atribuído

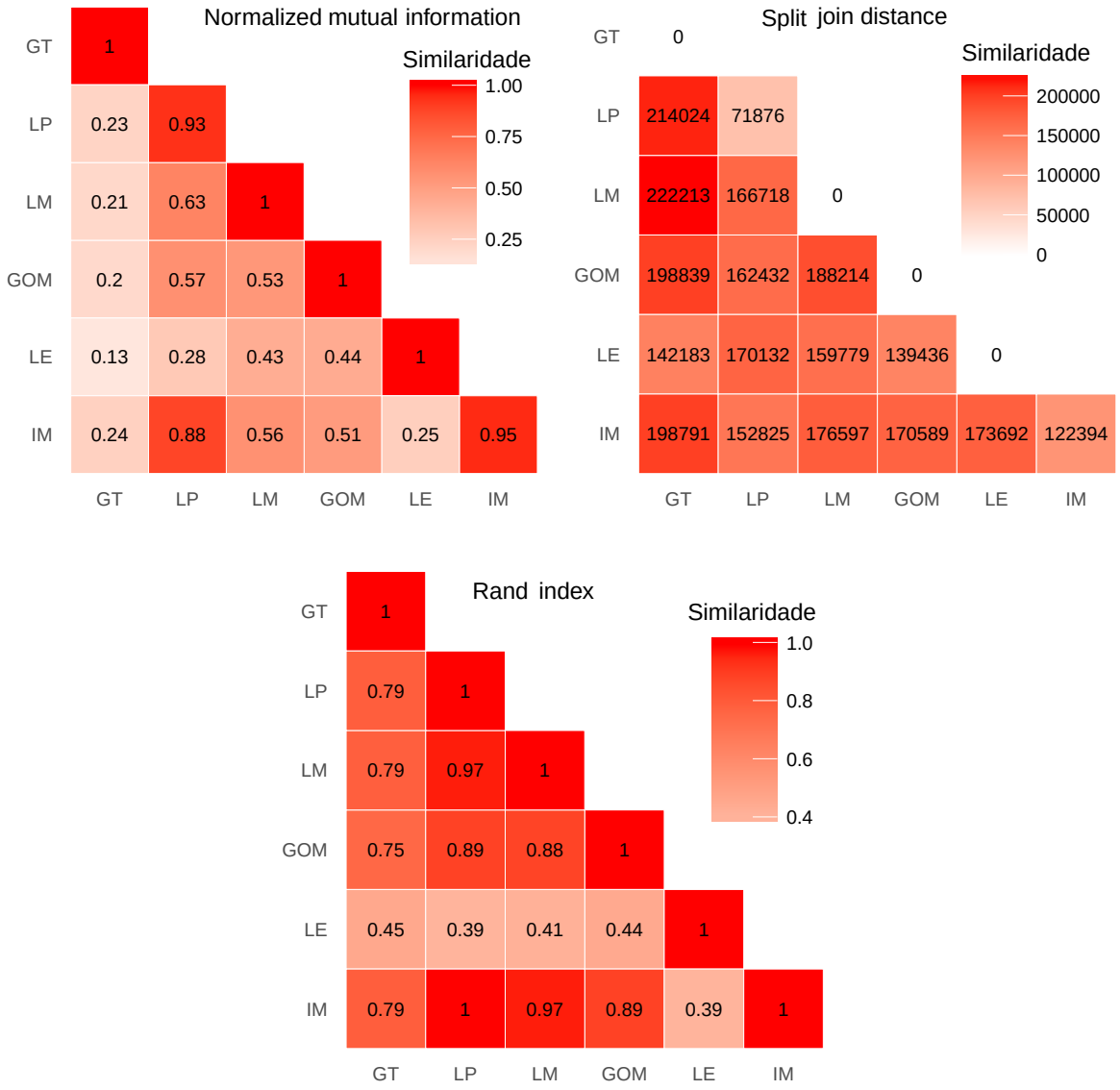

Figura 4.2: Similaridade entre o *ground truth* (GT) e as comunidades detectadas na rede APS Physics.

a cada um deles. Com base em um indicador topológico, em nosso caso calculado com a métrica de sobreposição de vizinhança (NO), rótulos são atribuídos a cada relacionamento conforme ilustrado na Figura 2.2. Assim, cada relacionamento entre pares de vértices é classificado como um laço forte quando o número de vizinhos em comum entre v_i e v_j é maior que um limiar. Nos demais casos, o relacionamento é classificado como um laço fraco.

Além do aspecto topológico, representado pela métrica NO, o modelo de força dos laços utilizado se baseia na persistência dos relacionamentos [Vaz de Melo et al., 2015].

Tabela 4.4: Classe atribuída a um relacionamento considerando o valor de cada aspecto. Fonte (adaptado): Vaz de Melo et al. [2015].

Classe do relacionamento	Indicador Topológico NO	Indicador Temporal P
Social	Forte	Persistente
	Fraco	Persistente
	Forte	Raro
Aleatório	Fraco	Raro

A persistência P de um relacionamento representa a regularidade das interações entre pares de vértices v_i e v_j. Essa regularidade é dada pela soma dos intervalos de tempo em que ocorre alguma interação entre os pares, conforme detalhado pela Equação 4.1,

$$P(v_i, v_j) = \frac{1}{t} \sum_{\kappa=1}^{t} [(v_i, v_j) \in \varepsilon_\kappa] \tag{4.1}$$

onde ε_κ representa os pares que interagiram durante o intervalo de tempo κ. Por ser uma métrica relacionada ao tempo dedicado a um relacionamento, a persistência é considera uma boa variável indicadora da força dos relacionamentos [Marsden & Campbell, 1984]. Nesta dissertação, usamos a métrica de persistência e também a sobreposição de vizinhança para medir a força dos relacionamentos através do algoritmo *RECAST*.

Assim, o histórico de interações entre pares de vértices v_i e v_j permite explorar o aspecto temporal para determinar, a partir de um limiar pré-definido, se o relacionamento é raro ou persistente. Então, com base na força e na persistência de um relacionamento, ele pode ser classificado como social ou aleatório [Vaz de Melo et al., 2015], conforme apresentado na Tabela 4.4.

Para o propósito desta dissertação, identificamos os relacionamentos *aleatórios* e *sociais* usando o classificador de relacionamentos *RECAST* (*Random Relationship Classifier Strategy*) [Vaz de Melo et al., 2015]. O *RECAST* classifica relacionamentos atribuindo um rótulo a cada par de vértices que possuem alguma aresta de interação na rede de agregação temporal. Assim, a partir de um conjunto de interações temporais, esse classificador é capaz de caracterizar relacionamentos aleatórios além de identificar diferentes tipos de relacionamento social (amizade, ponte e conhecido).

Modelo Nulo. Também utilizamos na etapa de avaliação um método estocástico de classificação em que cada aresta tem a mesma chance de ser rotulada como aleatória ou social. Esse método nos permitiu utilizar o arcabouço para comparar a filtragem baseada no *RECAST* com um modelo nulo de filtragem de relacionamentos[5]. Dado

[5]No modelo nulo a força dos relacionamentos é definida por uma variável aleatória.

Tabela 4.5: Percentual de mudança nas métricas sobre as redes sociais.

Rede	I	$\|V\|$	$\|E\|$	$\|E_R\|$	Δ	α	D	d	CC
APS	1^{st}	82	53	47	31	64	77	200	191
PubMed	5^{th}	91	46	54	23	50	54	150	173
DBLP	3^{rd}	61	38	62	15	63	104	146	381
Dartmouth	5^{th}	76	13	87	15	17	22	350	102
USC	10^{th}	03	02	98	14	81	2919	65	433
Enron	3^{rd}	12	01	99	07	06	55	150	92

I: iteração na qual foi obtido o percentual apresentado a partir do valor da rede original. V: conjunto de vértices; E: conjunto de arestas; E_R: conjunto de arestas aleatórias; Δ: grau máximo; α: grau médio; D: densidade; d: diâmetro; CC: coeficiente de agrupamento.

um grafo simples G_{os}, que representa uma rede com m relacionamentos, e um número k, cada relacionamento r_i é removido da rede com probabilidade p até alcançar o limite de k relacionamentos removidos. Ao final, todos os vértices desconectados também são removidos de G_{os}. Sobre a rede resultante S_a, avaliamos a qualidade das comunidades detectadas assim como feito sobre a rede S, filtrada a partir da classificação realizada pelo *RECAST*. Note que, a partir de uma sequência de interações, obtivemos as redes filtradas S e S_a pela remoção do mesmo número k de relacionamentos. Assim, pudemos verificar se o classificador utilizado em nosso arcabouço, possui uma probabilidade de acerto (verdeiro positivo) maior que a chance aleatória.

4.3 Melhoria na Detecção de Comunidades

Nesta seção apresentamos os conjuntos de evidências sobre o ganho de qualidade nas comunidades detectadas após as redes serem filtradas pelo nosso arcabouço. Como já mencionado, a diversidade de definições de comunidade e de redes de domínios distintos exige que a avaliação dessa melhoria seja feita pelo cruzamento de múltiplas estratégias a fim de identificar e reduzir o viés de alguma estratégia, métrica ou conjunto de dados. Assim, reunimos as evidências obtidas por diferentes métricas em linhas que esclarecem aspectos próprios da qualidade de uma comunidade que são aqueles de caráter estrutural e funcional, além da avaliação por comparação com *baselines*.

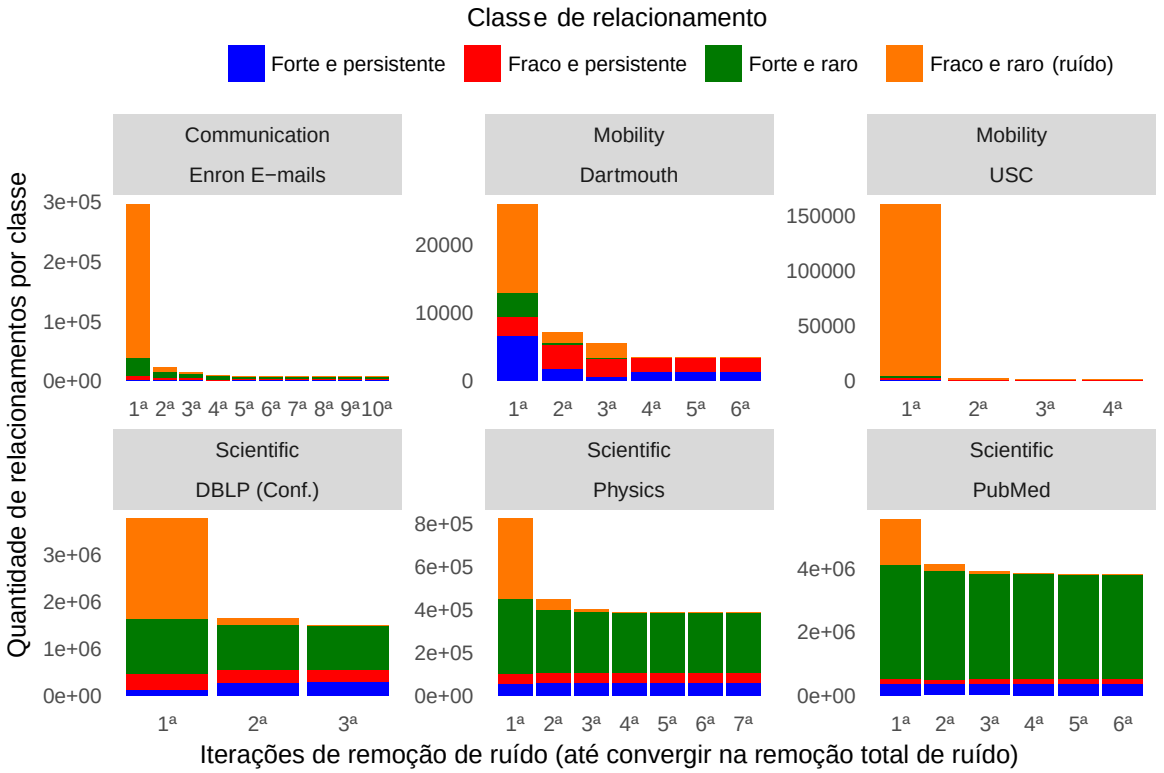

Figura 4.3: Classes de relacionamento ao final de cada iteração de convergência do processo de filtragem de relacionamentos sociais.

4.3.1 Evidências Estruturais

Para uma determinada métrica, a razão entre o seu valor na rede filtrada e o seu valor na rede original representa a porcentagem de alteração em relação a essa métrica. A Tabela 4.5 revela essa porcentagem de mudanças para diferentes métricas topológicas nas redes consideradas, o que indica o quanto o ruído interfere na caracterização de sua estrutura.

Ao analisar a Figura 4.3 é possível distinguir cada rede social pela quantidade total de relacionamentos aleatórios. Assim, as redes de mobilidade e comunicação são aquelas com maior proporção de relacionamentos aleatórios. Além disso, a maioria de seus relacionamentos são classificados como aleatórios (veja a coluna E_R na Tabela 4.5). Como consequência, um número maior de vértices que possuem todos os seus relacionamentos aleatórios são desconectados dessas redes porque não possuem participação distinta e significante em uma comunidade específica.

Conforme mostrado na Figura 4.3, para a maioria das redes, os relacionamentos aleatórios são removidos com poucas iterações até a convergência de remoção total de arestas que causam ruído e, consequentemente, vértices aleatórios. O grau máximo

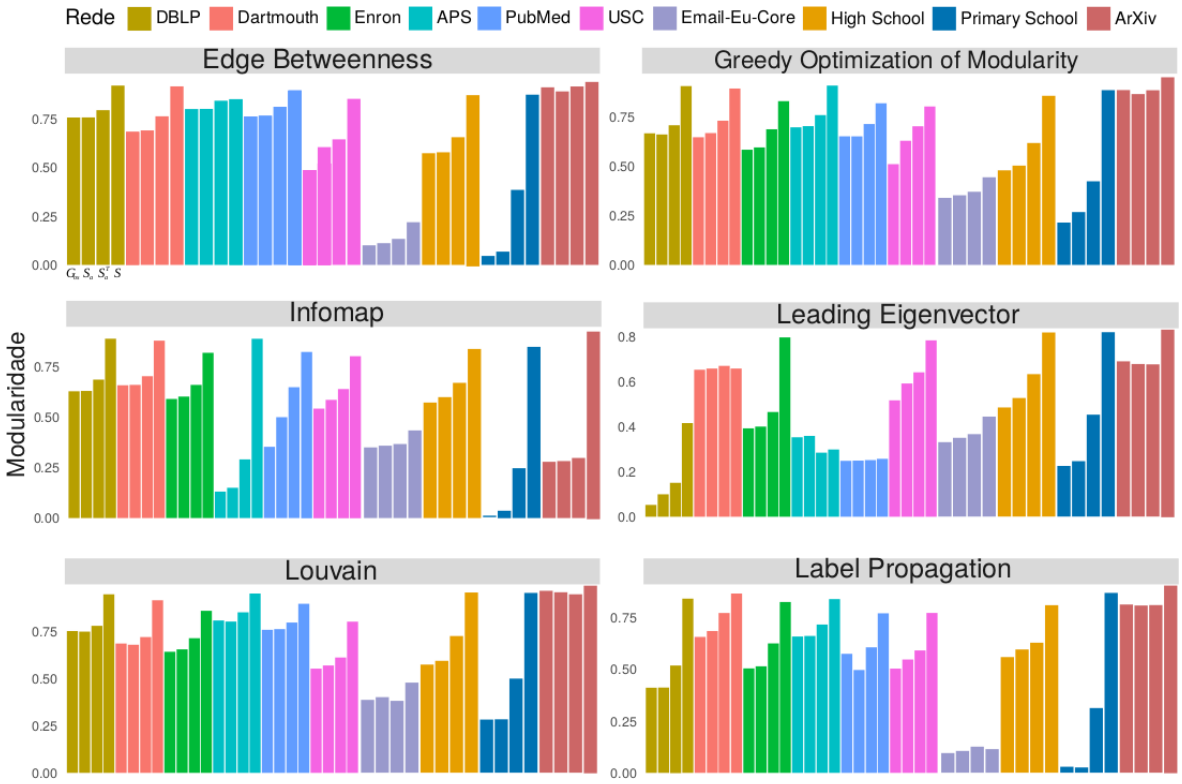

Figura 4.4: Modularidade das comunidades em cada rede (cor distinta) e para diferentes versões da mesma rede (mesma cor no eixo x), nesta ordem: G_{os}, S_a (modelo nulo), S_a^T (filtrada pela remoção de vértices aleatórios) e S.

das redes sociais também é afetado na proporção da quantidade de ruído removido. Além disso, o grau dos *hubs* foi reduzido em proporções maiores do que dos demais vértices da rede. Em outras palavras, a partir de um limiar de tamanho, o conjunto de relacionamentos de cada membro da rede tende a ser constituído predominantemente por interações aleatórias o que pode ser explicado pelo limite natural que os membros da rede têm em gerenciar seus contatos [David et al., 2010; Dunbar, 1992].

Outra propriedade topológica que se tornou mais explícita nas redes após a remoção do ruído é a tendência em formar grupos, que foi medida pelo coeficiente de agrupamento. Conforme o percentual de alteração mostrado na Tabela 4.5, o coeficiente de agrupamento global (CC) da rede filtrada se manteve igual ou maior que o da rede original, mesmo para as redes que se tornaram mais esparsas após a remoção de ruído[6]. Dentre as alterações observadas nas redes, as principais melhorias puderam ser verificadas sobre as suas estruturas de comunidade.

Assim, durante a remoção dos relacionamentos aleatórios, em cada iteração (Figura 4.3), há um aumento significativo na qualidade das comunidades detectadas pelos

[6]O esperado é que redes reais se apresentem esparsas e o número de arestas aumente linearmente com o número de vértices [Barabási, 2014].

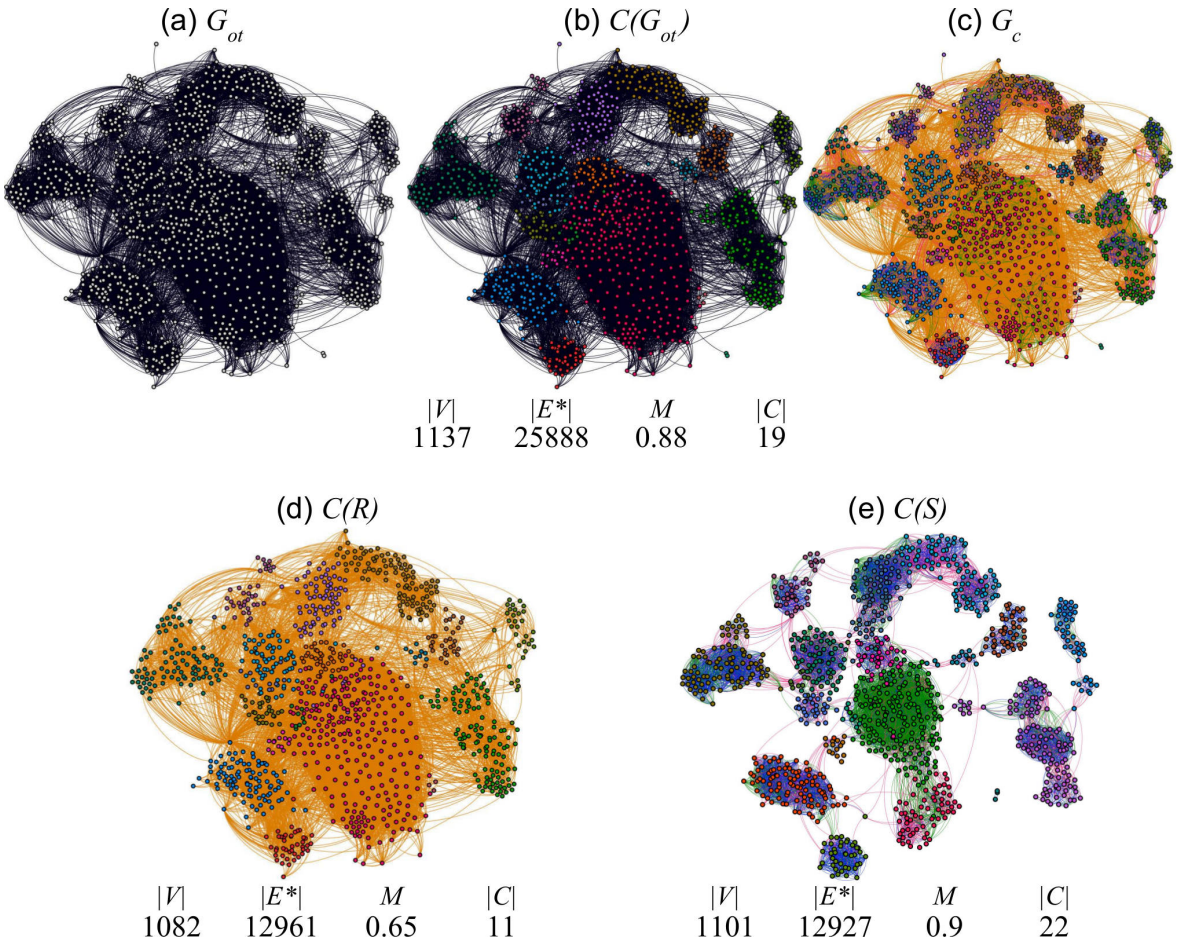

Figura 4.5: Conjunto de vértices (V), arestas ($E*$ - representa relacionamentos) e comunidades (C) e a medida de modularidade (M) da rede *Dartmouth* em diferentes etapas que compõem uma iteração do filtro de relacionamentos sociais: (a) rede original; (b) comunidades detectadas pelo algoritmo *Louvain*; (c) arestas classificadas pelo *RECAST*; (d) rede residual (R); (e) apenas os relacionamentos sociais da rede (S). A cor do vértice representa a comunidade a que um vértice pertence e a cor da aresta é baseada nas classes de relacionamento ou preto, antes de classificar.

algoritmos em cada rede. Essa melhoria foi verificada inicialmente pelo aumento na modularidade da rede ao ser filtrada, como mostrado na Figura 4.4. Por outro lado, a estrutura da rede residual R, formada apenas por relacionamentos aleatórios, é fracamente modular. Os menores valores de modularidade e de diâmetro da rede R evidenciam a sua maior semelhança com o modelo aleatório de rede, como o *Erdos-Renyi* [Erdős & Rényi, 1960], quando comparados aos valores correspondentes da rede original e da rede filtrada.

As estruturas da rede filtrada S e da rede R são exemplificadas na Figura 4.5d para a rede *Dartmouth*. Essa rede é originalmente muito modular e obteve ganho em modularidade em apenas uma iteração da etapa de remoção de relacionamentos

Tabela 4.6: Alteração no número de comunidades das redes após a remoção do ruído e no desvio padrão desse número obtido por algoritmos distintos sobre uma mesma rede.

Rede	LP	LM	GOM	IM	LE	*DP*
APS	24K/17K	14K/5K	14K/7K	21K/73K	13K/5K	4K/26K
PubMed	49k/34k	20k/9k	21k/10k	35k/203k	19k/9k	12k/76k
DBLP	130K/60K	80K/30K	53K/37K	130K/352K	47K/28K	36K/126K
arXiv	5K/5K	4K/4K	4K/4K	4K/13K	4K/3K	146/4K
Dartmouth	45/22	26/11	23/10	52/29	26/12	11.8/7.5
USC	141/9	130/10	130/8	146/42	127/13	7/12
High S.	17/8	15/7	15/3	18/9	15/9	1/2
Primary S.	25/48	14/6	14/8	23/95	14/13	5/34
Enron	659/3K	559/2K	573/2K	767/3K	522/1K	87/683
E. Eu-core	3/1	9/8	10/9	25/23	12/8	7.3/7.2

DP: desvio padrão do número de comunidades detectadas na rede (filtrada/original).

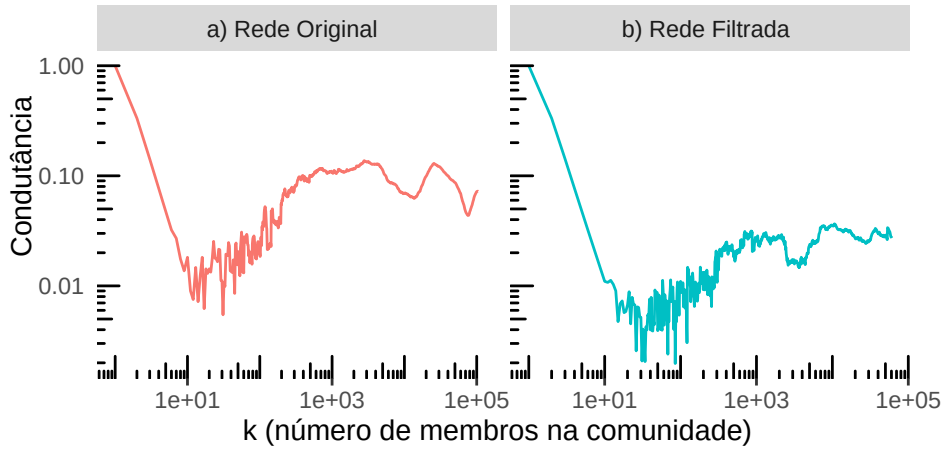

Figura 4.6: Condutância da rede APS para diferentes tamanhos de comunidade.

aleatórios. Embora esse ganho seja apenas de 2% (um dos menores dentre todas as redes), visualmente grupos coesos são mais claramente identificados na Figura 4.5e. Além disso, as comunidades detectadas pelo algoritmo *Louvain* na rede original $C(G_{ot})$ (Figura 4.5a) não são exatamente iguais às da rede filtrada $C(S)$ (Figura 4.5e). Ao comparar os valores de modularidade das redes, confirmamos que a rede se tornou mais modular depois de filtrada. Adicionalmente, a rede residual R tem modularidade muito menor do que a rede original.

Olhando para a rede filtrada S (Figura 4.5e), é possível identificar o surgimento de novas comunidades, a maioria delas isoladas, ou seja, correspondem a subgrafos desconexos. Em outros casos, um grupo de vértices é detectado como uma comunidade devido à presença de relacionamentos aleatórios. Depois de serem filtrados, alguns

desses grupos têm todas as suas arestas removidas porque são identificadas como relacionamentos aleatórios e, por conseguinte, seus vértices são desconectados. Neste caso, esse tipo de grupo é chamado de comunidade aleatória. Comunidades aleatórias são mais notáveis nas redes Enron, Dartmouth e USC, que também são as que possuem maior proporção de relacionamentos aleatórios (ver Tabela 4.5). Devido à remoção dessas comunidades, essas redes sofreram mudanças mais significativas em sua estrutura. Para as demais redes, após a filtragem de ruído é possível verificar o aumento do número total de comunidades detectadas e do consenso no número de comunidades detectadas por diferentes algoritmos. Esse consenso, medido pelo desvio padrão do número de comunidades, convergiu entre os algoritmos, como mostrado na Tabela 4.6.

Além da redução no número de vértices apresentada para cada rede na Tabela 4.5, o aumento no número de novas comunidades (veja Tabela 4.6) contribuiu para uma redução em seu tamanho médio. Essa observação também evidencia o ganho em qualidade nas comunidades da rede filtrada conforme Figura 4.6, onde o *Network Community Profile* (NCP) caracteriza a melhor escala de tamanho de comunidade em um intervalo de possibilidades [Leskovec et al., 2008]. Além disso, a comparação com o NCP da rede original permite observar (Figura 4.6) que a rede filtrada possui melhor condutância (valor mais baixo) em todo o intervalo de tamanhos de comunidade possíveis.

Como já mencionado, a métrica de condutância tem uma tendência para dar melhores pontuações para agrupamentos com um número menor de grupos (porque mais grupos provavelmente terão mais arestas de corte) [Almeida et al., 2012; Zaki & Wagner Meira, 2014]. Mesmo com o aumento no número de comunidades nas redes, é obtida melhoria na condutância após a filtragem de relacionamentos aleatórios. Diferentes ganhos em melhoria foram verificados em todas as redes e pode ser observado no exemplo da rede APS apresentado na Figura 4.6. Também pode ser observado que o tamanho ideal das comunidades que otimiza a condutância é menor na rede filtrada e, portanto, mais próximo do tamanho típico de comunidades reais[7] [Leskovec et al., 2008]. Isso também reforça a percepção de que a estrutura de comunidade em um nível mais granular representa melhor a estrutura da rede sem ruído.

Note que mencionamos também que as métricas condutância e modularidade possuem um certo viés estrutural, gerando melhores resultados para um número menor de *clusters* [Almeida et al., 2012]. Em nosso caso, a obtenção de melhorias para ambas as métricas ocorre em um contexto oposto ao que leva ao viés porque houve um aumento no número de comunidades depois que a rede foi filtrada.

[7]As comunidades de uma rede real tendem a existir apenas em escalas de tamanho pequeno de até cerca de 100 vértices, enquanto em grandes escalas de tamanho as comunidades da rede se tornam menos comuns [Leskovec et al., 2008]

Tabela 4.7: Comparação entre técnicas de detecção de comunidade na rede APS.

Métrica/Alg.	LE	LM	LP	GOM	IM	WT	Média
NMI	61.5	4.7	8.7	10.0	4.2	8.7	16.3
RI	33.3	-1.3	-1.3	4.0	-1.3	0	5.6
SJD	29.4	15.6	36.1	8.4	12.6	30.8	22.2

Os números representam a porcentagem de ganho em qualidade nas comunidades detectadas quando comparadas ao *ground truth*.

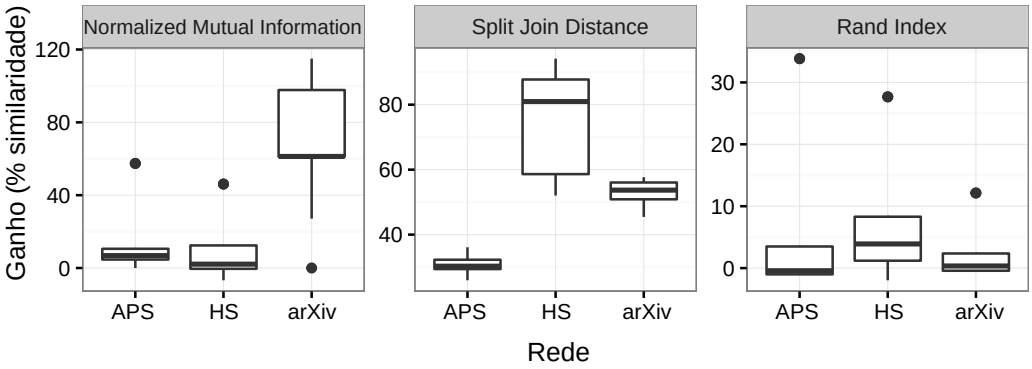

Figura 4.7: Box plot do percentual de ganho em similaridade entre comunidades detectadas e comunidades funcionais nas redes reais que possuem *ground truth*.

4.3.2 Evidências Funcionais

Além da utilização de métricas de qualidade estrutural, avaliamos as melhorias das comunidades detectadas através de métricas de similaridade. Essa avaliação envolveu a comparação da similaridade entre as comunidades da rede original $C(G_{ot})$, as comunidades da rede filtrada $C(S)$ e as comunidades funcionais $P(G)$.

Como verificado anteriormente, a rede APS apresentou originalmente alta sobreposição na participação dos pesquisadores em áreas distintas (comunidades funcionais, Tabela 4.2), baixa modularidade de suas comunidades funcionais e baixa semelhança entre essas comunidades e as comunidades detectadas. Apesar dessas características, melhorias significativas foram alcançadas em todos esses aspectos após a rede APS ser filtrada. A Figura 4.7 apresenta o ganho em similaridade entre comunidades estruturais e comunidades funcionais e, para a rede APS, esse ganho é considerável para a maioria dos algoritmos de detecção de comunidades, obtendo, em média, aumento entre 5% e 22% na similaridade, conforme detalhado na Tabela 4.7. Considerando todas as redes da Figura 4.7, o ganho médio foi de até 95% e máximo de 115%, obtido na rede arXiv.

As comunidades funcionais das redes arXiv e APS foram construídas a partir da área predominante em que cada pesquisador publica. Na Figura 4.7 é possível verificar,

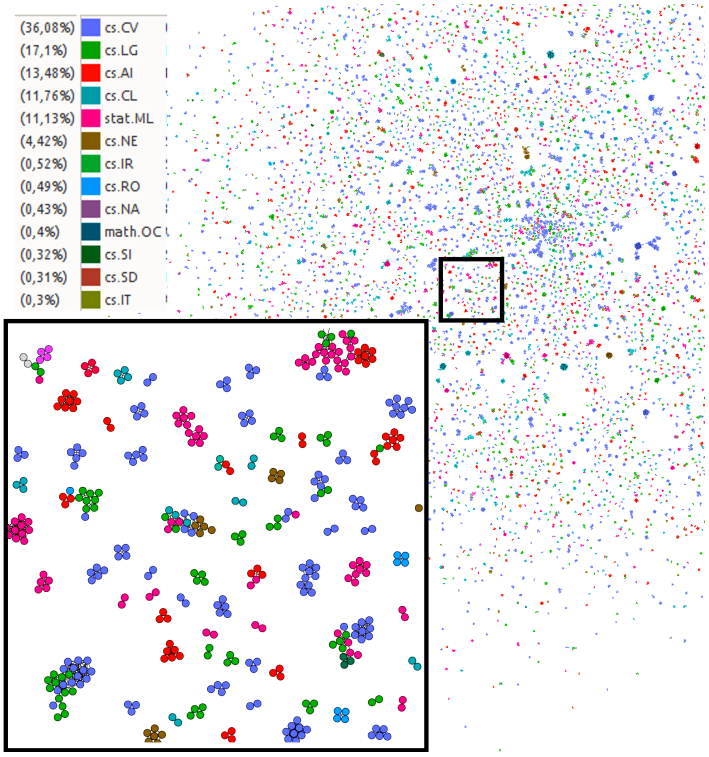

Figura 4.8: Layout estrutural da rede arXiv (*Force Atlas*). Cada cor nos vértices identifica a sua comunidade funcional (13 comunidades correspondem a 100% dos vértices) e é possível distinguir facilmente vértices que pertencem ao mesmo grupo coeso.

contudo, que a estimativa de ganho em similaridade da rede arXiv é consideravelmente maior que a estimativa para a rede APS. Isso ocorre entre essas redes em razão das diferenças em similaridade entre as suas comunidades funcionais e estruturais. Assim, na rede APS, essa similaridade é menor, pois um módulo geralmente contém muitos vértices de comunidades funcionais distintas. Essa análise comparativa foi feita entre as redes APS e arXiv em razão da disponibilidade de dados de *ground truth* nessas redes e por representarem, respectivamente, a rede com menor e uma das redes com maior similaridade entre suas comunidades funcionais e estruturais, dentre todas as redes para as quais possuímos o *ground truth*. Além de possuir maior similaridade entre as comunidades de sua rede original e as comunidades funcionais, a rede arXiv também obteve maior ganho nessa similaridade após a remoção de ruído, quando comparado ao ganho obtido na rede APS.

A partir dos valores baixos de similaridade inicial que são apresentados na Figura 4.9, é possível verificar que o menor ganho na rede APS é influenciado pela maior sobreposição entre suas comunidades funcionais, evidenciadas na Tabela 4.2. Essa característica de alta sobreposição da rede APS também é indicada pela baixa ho-

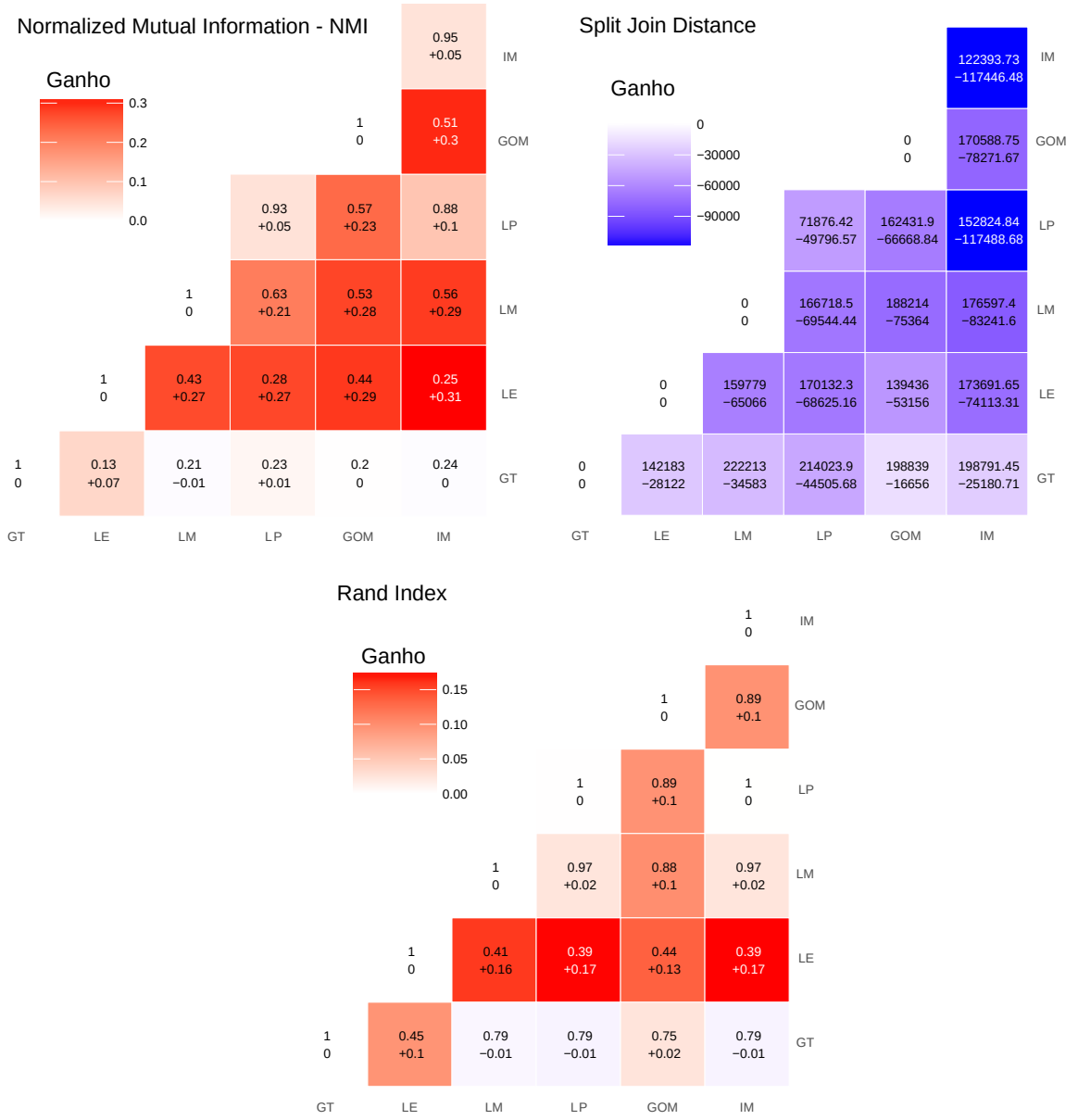

Figura 4.9: Ganho em similaridade (valor na parte superior) entre comunidades detectadas e o *ground truth* da rede APS. Na parte inferior de cada comparação, é exibido o valor de ganho para essa similaridade medido após a aplicação do filtro de relacionamentos sociais. Para as métricas *Normalized Mutual Information* e *Rand Index*, o ganho positivo significa aumento na similaridade. Por outro lado, na métrica *Split Join Distance* esse ganho em similaridade é representado por valores negativos.

mogeneidade de cores (comunidades funcionais) dentro das comunidades estruturais, conforme apresentado na Figura 4.1. Por outro lado, a Figura 4.8 permite visualizar que na rede arXiv muitas das comunidades estruturais (grupos coesos ou componentes) contêm vértices que pertencem a uma mesma comunidade funcional (uma cor).

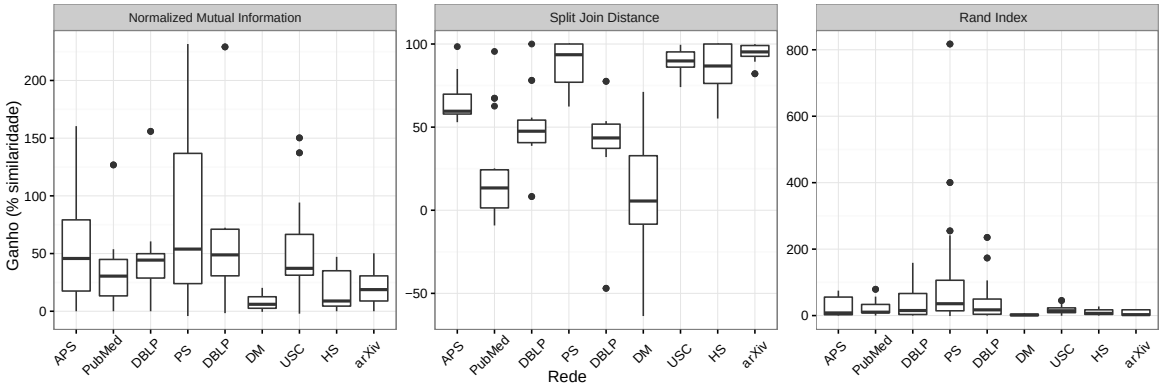

Figura 4.10: Box plot do percentual de ganho em similaridade entre as comunidades detectadas por diferentes algoritmos após a remoção de ruído.

Apesar de que a maioria dos algoritmos detectam comunidades com base em uma definição própria de comunidade [Cazabet et al., 2010; Coscia et al., 2011; Fortunato, 2010], ao remover o ruído conseguimos aumentar o consenso entre diferentes algoritmos sobre qual é a estrutura de comunidade que deve ser detectada. Esse consenso foi confirmado também pela redução da variância entre o número de comunidades detectadas por esses algoritmos. Essa variância reduziu em média para 38% da variância original, conforme detalhado na Tabela 4.6. Além disso, tornamos essas comunidades estruturais mais parecidas com as comunidades funcionais, o que verificamos com as três métricas de similaridade. Isto significa que o filtro de relacionamentos sociais permite a convergência entre diferentes definições estruturais e funcional de comunidade, como apresentado na Figura 4.10. Além disso, essa convergência permite estimar o quanto a aleatoriedade afeta o consenso entre a diversidade de abordagens de detecção ou definições da estrutura de comunidade.

4.3.3 Evidências Relativas a um *Baseline*

Na Figura 4.4 são apresentados, para cada algoritmo de detecção de comunidades, os valores de modularidade de cada rede social (identificada por uma cor). Assim, para uma mesma rede, a primeira barra indica a modularidade das comunidades detectadas sobre a rede original G_{os} e as barras seguintes sobre as redes filtradas por diferentes métodos. Na maior parte, esses valores são maiores nas redes S, filtradas pelo nosso arcabouço utilizando o *RECAST* (quarta barra). Em seguida, a segunda maior modularidade é mais observada para a remoção de vértices aleatórios (terceira barra) e por último para o modelo nulo (segunda barra) que possui valores aproximados aos da rede original.

Tabela 4.8: Percentual de evidências que indicaram melhoria considerável na detecção de comunidades por 6 algoritmos sobre 11 redes.

Avaliação:	Estrutural				Funcional	Relativa a um modelo	
Principal Métrica	Q	Φ	$C(G)$	DP	$P(G)$	Baseline	Nulo
Casos Favoráveis	97%	82%	91%	80%	89%	95%	98%

Q: modularidade; Φ: condutância; $C(G)$: similaridade entre comunidades detectadas; DP: desvio padrão do número de comunidades detectadas por algoritmos distintos; $P(G)$: similaridade entre comunidades detectadas e *ground truth*.

Além de evidenciar a eficácia da remoção de relacionamentos aleatórios, o estudo dos vértices aleatórios nos permitiu confirmar que, ao remover vértices com alta centralidade de grau (*hubs*), obtém-se ganho em modularidade das comunidades detectadas conforme demonstrado primeiramente por Wen et al. [2011]. Ademais, verificamos que o grau dos *hubs* é fortemente correlacionado com a alta proporção de relacionamentos aleatórios e, por isso, ao remover vértices violadores ou *hubs*, uma quantidade considerável das arestas aleatórias também é removida. Contudo, nota-se que geralmente tais vértices possuem relacionamentos que não são aleatórios e por isso pertencem a alguma comunidade, diferente do que ocorre com os vértices aleatórios. Além disso, a alta importância dos *hubs* na rede, medida pela centralidade de grau, não justifica a sua remoção. Dessa forma, a filtragem de uma rede por remoção de relacionamentos aleatórios se apresenta mais adequada que a remoção de vértices violadores.

4.3.4 Análise Geral das Evidências de Melhoria

Pelo uso de estratégias de avaliação baseadas em pressupostos distintos sobre a qualidade de uma comunidade, obtivemos as evidências descritas anteriormente. Neste contexto, a avaliação da qualidade de comunidades em redes que passam por um método de filtragem nos permite estimar o viés entre as estratégias utilizadas, em alguma das métricas consideradas ou em algum dos conjuntos de dados utilizados. Assim, a seguir são apresentados os resultados gerais da melhoria de qualidade das comunidades detectadas, separados por estratégia de avaliação.

Da rede resultante da filtragem de ruído, foram extraídas comunidades por diferentes técnicas de detecção. A qualidade dessas comunidades mostrou-se consideravelmente melhor que a da rede original na maioria das métricas de avaliação estrutural. Além disso, verificamos o aumento do consenso sobre as comunidades detectadas por diferentes técnicas. Na Tabela 4.8 é apresentado, separado por estratégia de avaliação, os percentuais de evidências que indicaram melhoria na detecção de comunidades.

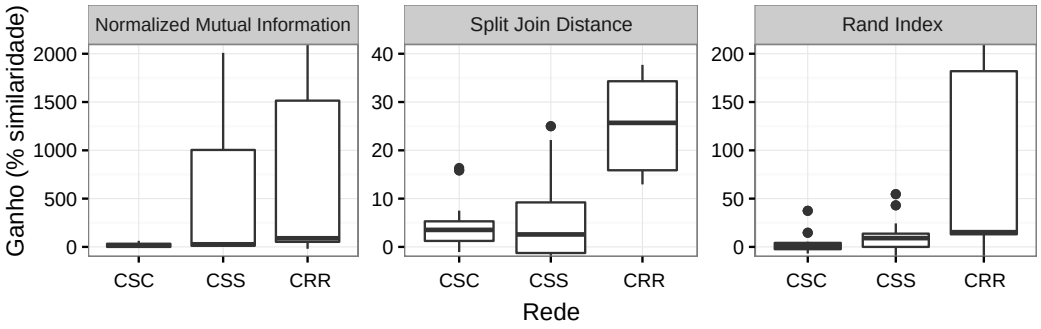

Figura 4.11: Box plot do percentual de ganho em consenso, ou seja, em similaridade entre as comunidades detectadas por diferentes algoritmos após a remoção de ruído da rede simulada com diferentes configurações do modelo.

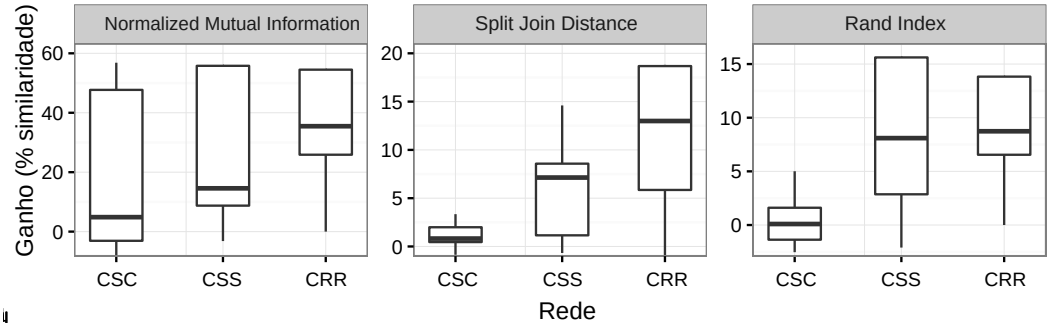

Figura 4.12: Box plot do percentual de ganho em similaridade entre as comunidades detectadas pelos seis algorítmos e as comunidades funcionais (*ground truths*), após a remoção de ruído das redes simuladas com diferentes configurações do modelo.

As comunidades da rede APS obtiveram o menor ganho em qualidade na avaliação funcional em relação às demais estratégias de avaliação. Através de uma análise das características dessas comunidades, também foi possível verificar que o menor ganho na avaliação é devido à menor similaridade entre as suas comunidades funcionais e estruturais. Por outro lado, nas demais redes que também possuem dados de *ground truth*, o ganho na avaliação funcional foi equiparado ao ganho obtido na avaliação estrutural. Consensualmente, em todas as redes e com valores consideravelmente diferentes, as três métricas de similaridade evidenciaram que as comunidades estruturais se tornaram mais parecidas com as comunidades funcionais após a remoção do ruído.

Ao ser comparada com outros métodos de filtragem como o modelo nulo, a filtragem de relacionamentos aleatórios apresentou resultados consideravelmente melhores para a mesma configuração de quantidade de arestas a serem removidas. Em comparação com a remoção de vértices violadores, os métodos e modelos utilizados em nosso arcabouço também foram mais adequados.

O uso de redes simuladas permitiu complementar as estratégias funcional e estrutural de avaliação de qualidade do arcabouço proposto. Além disso, o resultado dessa avaliação confirmou uma melhoria considerável na detecção de comunidades em duas das três configurações do gerador de modelos utilizado. Dentre essas redes simuladas, as redes de interações geradas a partir de uma rede estática real (com ou sem peso) obtiveram resultados significativamente melhores na maioria das estratégias de avaliação, como na avaliação estrutural (Figura 4.11) e funcional (Figura 4.12). De modo geral, essas melhorias se apresentaram em escalas de valores adequadas às redes reais, demonstrando que o gerador utilizado permite a geração de redes sintéticas com *ground truths* e modelagem realística das sequências de interações.

Diante da discrepância entre os ganhos obtidos por diferentes estratégias de avaliação, foi possível verificar que o uso de apenas uma métrica de qualidade estrutural ou funcional torna os resultados pouco confiáveis. Além disso, as múltiplas estratégias de avaliação explicitam valores extremos em algumas das métricas. Por exemplo, em nossos resultados, pudemos verificar o viés da métrica de modularidade ao utilizar o método estocástico de filtragem para remover uma quantidade de arestas maior que outros métodos. Neste contexto, também foi possível verificar a inadequação da granularidade das comunidades funcionais na rede APS. Portanto, a obtenção de consenso entre diferentes estratégias foi fundamental para uma avaliação robusta sobre a melhoria da qualidade na tarefa de detecção de comunidades.

Capítulo 5

Conclusões e Trabalhos Futuros

Uma das principais contribuições desta dissertação é um arcabouço de filtragem de relacionamentos sociais que, através do algoritmo *RECAST*, considera os aspectos topológicos e temporais de redes sociais para melhorar a detecção de comunidades em redes estáticas. Esses aspectos caracterizam um modelo de força dos laços que foi utilizado em nosso arcabouço para remover relacionamentos aleatórios, ou seja, que ocorrem entre pares de indivíduos com pouca probabilidade de interagir novamente. Após a filtragem desses relacionamentos, construímos uma rede estática composta apenas por relacionamentos sociais.

Especificamente, a partir de uma sequência de interações temporais, utilizamos o algoritmo *RECAST* para classificar os relacionamentos como sociais ou aleatórios, com base nos seus valores de persistência e de sobreposição de vizinhança. Assim, iterativamente classificamos e removemos todos os relacionamentos com alta probabilidade de serem aleatórios. Ao final das iterações de filtragem, obtivemos uma rede livre de ruído sobre a qual analisamos a qualidade da estrutura de comunidade utilizando diferentes estratégias de avaliação.

Os experimentos realizados envolveram dez redes sociais reais de domínios distintos e três configurações de modelos de redes simuladas. Sobre esse conjunto de dados, avaliamos o nosso arcabouço comparando-o com um método de filtragem de rede proposto na literatura e um método estocástico de remoção de arestas. Adicionalmente, a partir desse conjunto de experimentos, também avaliamos a qualidade das comunidades detectadas por diferentes técnicas de detecção, utilizando diversas instâncias e versões de redes com e sem ruído. Nesse contexto, a avaliação evidenciou uma clara melhoria na qualidade dessas comunidades em mais de 80% dos casos.

Nesse processo, conseguimos responder às nossas questões de pesquisa. Primeiro, comprovamos que relacionamentos aleatórios causam ruído na rede e que esse ruído tem

efeito negativo na qualidade de comunidades detectadas pelos algoritmos utilizados. Especificamente, verificamos que a presença de relacionamentos aleatórios aumenta o erro na associação de vértices a comunidades por algoritmos existentes. A segunda resposta esclarece que a avaliação da qualidade das comunidades detectadas a partir de uma rede filtrada deve ser feita pela coleta de múltiplas evidências. E mais importante, no cenário formado pelos conjuntos de dados que utilizamos, essas evidências devem compor pelo menos três estratégias que permitam avaliar definições distintas de comunidade. Por sua vez, cada estratégia deve ser composta por métricas independentes e preferencialmente divergentes para que possam medir aspectos distintos da qualidade de uma comunidade. Por fim, esclarecemos que, em resposta à nossa última questão de pesquisa, a filtragem de relacionamentos aleatórios é consensualmente benéfica e na maioria dos casos (entre 80% e 98%) foram observados ganhos consideráveis. Além disso, verificamos que o tipo de interação das redes utilizadas interfere na qualidade dos resultados mais que a técnica utilizada.

Nosso arcabouço permite o acoplamento de diferentes modelos de força dos laços para classificar os relacionamentos e a escolha de qual das classes de relacionamento disponíveis no modelo devem ser filtradas. Entretanto, uma limitação do nosso trabalho é o uso de um único modelo de força dos laços que, mesmo assim, permitiu atingir satisfatoriamente os objetivos propostos, particularmente em relação à melhoria na detecção de comunidades com uso de propriedades mínimas das redes sociais. Apesar disso, outros atributos podem ser explorados considerando a disponibilidade no domínio de aplicação, como o peso ou outros atributos dos vértices e das arestas. Com isso, pode ser experimentado um nível de especialização maior sobre o domínio da rede social e, consequentemente, um refinamento do modelo de força dos laços. Nesse caso, o uso do arcabouço proposto com outros modelos é capaz de permitir resultados ainda melhores e que não foram objetivo desta dissertação.

Por exemplo, o modelo de força dos laços utilizado pelo classificador proposto por Brandão et al. [2017] pode ser experimentado. Naturalmente que, em outros classificadores, o ruído corresponde às classes de relacionamento que representam a maior probabilidade do relacionamento ser gerado aleatoriamente. Assim, a qualidade da modelagem do ruído depende da qualidade das variáveis escolhidas para medir a força dos laços que, por sua vez, dependem da capacidade de representar as características dos relacionamentos sociais específicas a um domínio.

Nesse sentido, pretendemos em trabalhos futuros refinar a definição de ruído e apresentar os resultados do uso do arcabouço proposto com um modelo alternativo de força dos laços baseado na duração das interações. Além disso, pretendemos avaliar o uso de outras variáveis, como a *recência* que pode ser considerada para classificar os

relacionamentos como recentes ou antigos com base na data das interações dos vértices envolvidos.

Considerando que o arcabouço proposto permite mapear sequências de interações sociais em redes estáticas filtradas, a qualidade da representação dessas redes sociais pode ser deduzida e estimada, permitindo-nos especificar o método, modelos e parâmetros utilizados para construí-las. Também é possível determinar a quantidade de ruído removido e o ganho em qualidade obtido. Dessa forma, pretendemos usar esse arcabouço de filtragem para construir e especificar conjuntos de dados adicionais em nosso repositório[1].

Em nossos experimentos, a aplicação do arcabouço de filtragem convergiu para a remoção total do ruído das redes utilizadas. Foi possível quantificar o ruído em cada uma das redes e caracterizá-las pela proporção de aleatoriedade em seus conjuntos de relacionamentos. Também foi possível distinguir redes de diferentes domínios apenas pela proporção de seus relacionamentos aleatórios. Por exemplo, verificamos que na troca de *e-mails* na rede Enron e nos contatos nas redes de mobilidade acadêmica, a aleatoriedade é geralmente maior do que nas redes de coautoria científica.

O uso do nosso arcabouço configurado para remover ruído de redes sociais mostrou-se como uma das nossas mais promissoras contribuições para a tarefa de detecção de comunidades. Além disso, pretendemos apresentar em trabalhos futuros os resultados obtidos para outras tarefas e com diferentes configurações, como na filtragem de outras classes de relacionamento além da *aleatória*. Por exemplo, pretendemos demonstrar a obtenção de outras representações de redes sociais, como a da estrutura de *backbone*[2]. Ademais, as redes filtradas por nosso arcabouço poderão ser avaliadas quanto à melhoria de qualidade na detecção de comunidades temporais ou definidas a partir de outras características como a sua dinâmica ou a sobreposição entre elas.

Espera-se que novos métodos de detecção de comunidade e análise de redes sociais possam fazer proveito dos resultados apresentados nesta dissertação, em especial a influência do aspecto temporal, o mapeamento da sequência de interações em redes estáticas após a filtragem da rede e as estratégias de avaliação propostas. Grande parte do que foi revelado nesta dissertação se baseia na combinação de conceitos e na confirmação de teorias propostas anteriormente como a força dos laços e a estrutura de comunidade. Contudo, o contexto de aplicação e a proposta metodológica compreendem contribuições novas, além de que foram explicitados detalhes que não estão presentes nesses relatos e utilizados conjuntos de dados de larga escala.

---

[1] http://cnet.jcloud.net.br/

[2] A extração de *backbone* consiste no isolamento das estruturas relevantes para representação reduzida, porém significativa, da rede [Serrano et al., 2009].

Referências Bibliográficas

Abrahao, B.; Soundarajan, S.; Hopcroft, J. & Kleinberg, R. (2012). On the separability of structural classes of communities. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 624–632, New York, NY, USA. ACM.

Abufouda, M. & Zweig, K. A. (2015). Are we really friends?: Link assessment in social networks using multiple associated interaction networks. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 771–776, New York, NY, USA. ACM.

Abufouda, M. & Zweig, K. A. (2017). Link classification and tie strength ranking in online social networks with exogenous interaction networks. *CoRR*, abs/1708.04030.

Adamic, L. A. & Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, 25(3):211–230.

Almeida, H.; Guedes, D.; Meira, W. & Zaki, M. (2011). Is there a best quality metric for graph clusters? *Machine Learning and Knowledge Discovery in Databases*, pp. 44–59.

Almeida, H.; Guedes, D.; Meira Jr, W. & Zaki, M. J. (2012). Towards a Better Quality Metric for Graph Cluster Evaluation. *Journal of Information and Data Management*, 3(3):378–393.

Alves, B. L. (2013). *Um Estudo sobre a Evolução Temporal de Comunidades Científicas*. Dissertação de Mestrado, Universidade Federal de Minas Gerais.

Alves, B. L.; Benevenuto, F. & Laender, A. H. (2013). The Role of Research Leaders on the Evolution of Scientific Communities. In *Proceedings of the 22nd International Conference on World Wide Web, Companion Volume*, pp. 649–656, New York, NY, USA. ACM.

Aral, S. & Walker, D. (2014). Tie strength, embeddedness, and social influence: A large-scale networked experiment. *Management Science*, 60(6):1352–1370.

Barabási, A. L. (2014). *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*. Basic Books.

Barabâsi, A.-L.; Jeong, H.; Néda, Z.; Ravasz, E.; Schubert, A. & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3):590–614.

Barabási, A.-L. & Pósfai, M. (2016). *Network science*. Cambridge University Press, Cambridge.

Barber, M. J. & Clark, J. W. (2009). Detecting network communities by propagating labels under constraints. *Phys. Rev. E*, 80:026129.

Barrat, A.; Barthlemy, M. & Vespignani, A. (2008). *Dynamical Processes on Complex Networks*. Cambridge University Press, New York, NY, USA, 1st edição.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Brandão, M. A. & Moro, M. M. (2017). The strength of co-authorship ties through different topological properties. *Journal of the Brazilian Computer Society*, 23(1):5.

Brandão, M. A.; Vaz de Melo, P. O. S. & Moro, M. M. (2017). STACY: Um Novo Algoritmo para Automaticamente Classificar a Força dos Relacionamentos ao Longo dos Anos. In *Anais do XXXII Simpósio Brasileiro de Bancos de Dados, Uberlândia, MG, Brazil, October 4-7, 2017.*, pp. 136–147.

Burt, R. S. (1992). Structural holes the social structure of competition. *Explorations in economic sociology*, 65:103.

Casteigts, A.; Flocchini, P.; Quattrociocchi, W. & Santoro, N. (2011). *Time-Varying Graphs and Dynamic Networks*, pp. 346–359. Springer Berlin Heidelberg, Berlin, Heidelberg.

Cazabet, R. & Amblard, F. (2014). *Dynamic Community Detection*, pp. 404–414. Springer New York, New York, NY.

Cazabet, R.; Amblard, F. & Hanachi, C. (2010). In *IEEE Second International Conference on Social Computing, title=Detection of Overlapping Communities in Dynamical Social Networks*, pp. 309–314.

Clauset, A. (2005). Finding local community structure in networks. *Phys. Rev. E*, 72(2):26132.

Clauset, A.; Newman, M. E. J. & Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E*, 70:066111.

Coscia, M.; Giannotti, F. & Pedreschi, D. (2011). A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(5):512–546.

Danon, L.; Díaz-Guilera, A.; Duch, J. & Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008.

David, E.; Jon, K.; Easley, D. & Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA.

De Domenico, M.; Lancichinetti, A.; Arenas, A. & Rosvall, M. (2015). Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Phys. Rev. X*, 5:011027.

Dongen, S. V. (2000). Performance criteria for graph clustering and Markov cluster experiments. Relatório técnico, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, The Netherlands, The Netherlands.

Dunbar, R. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469 – 493.

Erdős, P. & Rényi, A. (1960). On the evolution of random graphs. In *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, pp. 17–61.

Euler, L. (1736). Solutio problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 8:128–140.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3–5):75–174.

Fortunato, S. & Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41.

Fortunato, S. & Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659:1–44.

Gemmetto, V.; Barrat, A. & Cattuto, C. (2014). Mitigation of infectious disease at school: targeted class closure vs school closure. *BMC Infectious Diseases*, 14(1):695.

Gilbert, E. & Karahalios, K. (2009). Predicting Tie Strength with Social Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 211–220, New York, NY, USA. ACM.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380.

Greene, D.; Doyle, D. & Cunningham, P. (2010). Tracking the Evolution of Communities in Dynamic Social Networks. In *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining*, pp. 176–183.

He, J. & Chen, D. (2015). A fast algorithm for community detection in temporal network. *Physica A: Statistical Mechanics and its Applications*, 429(Supplement C):87–94.

Holme, P. (2015). Modern temporal network theory: a colloquium. *The European Physical Journal B*, 88(9):234.

Holme, P. & Saramäki, J. (2012). Temporal networks. *Physics reports*, 519(3):97–125.

Holme, P. & Saramäki, J. (2013). *Temporal Networks*. Understanding Complex Systems. Springer Berlin Heidelberg.

Hric, D.; Darst, R. K. & Fortunato, S. (2014). Community detection in networks: Structural communities versus ground truth. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.*, 90(6):62805.

Hric, D.; Peixoto, T. P. & Fortunato, S. (2016). Network structure, metadata, and the prediction of missing nodes and annotations. *Phys. Rev. X*, 6:031038.

Kivelä, M.; Arenas, A.; Barthelemy, M.; Gleeson, J. P.; Moreno, Y. & Porter, M. A. (2014). Multilayer networks. *Journal of complex networks*, 2(3):203–271.

Klymko, C.; Gleich, D. & Kolda, T. G. (2014). Using triangles to improve community detection in directed networks. *arXiv preprint arXiv:1404.5874*.

Kossinets, G. & Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science*, 311(5757):88–90.

Kostakos, V. (2009). Temporal graphs. *Physica A: Statistical Mechanics and its Applications*, 388(6):1007–1023.

Kumpula, J. M. & Kaski, K. (2008). A sequential algorithm for fast clique percolation. *Physical Review E*, 78(2):1–8.

Lambiotte, R.; Delvenne, J.-C. & Barahona, M. (2008). Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770*.

Lancichinetti, A.; Fortunato, S. & Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015.

Lazarsfeld, P. F.; Merton, R. K. et al. (1954). Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, 18(1):18–66.

Leskovec, J.; Kleinberg, J. & Faloutsos, C. (2007). Graph Evolution: Densification and Shrinking Diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1):2.

Leskovec, J.; Lang, K. J.; Dasgupta, A. & Mahoney, M. W. (2008). Statistical Properties of Community Structure in Large Social and Information Networks. In *Proceedings of the 17th International Conference on World Wide Web*, pp. 695–704, New York, NY, USA. ACM.

Leão, J. C.; Brandão, M. A.; Vaz de Melo, P. O. S. & Laender, A. H. F. (2017a). Classificação de Relações Sociais para Melhorar a Detecção de Comunidades. In Proceedings *of the VI Brazilian Workshop on Social Network Analysis and Mining*, São Paulo, SP, Brazil.

Leão, J. C.; Brandão, M. A.; Vaz de Melo, P. O. S. & Laender, A. H. F. (2017b). Mineração de Perfis Sociais em Redes Temporais. In *Anais do XXXII Simpósio Brasileiro de Bancos de Dados, Uberlândia, MG, Brazil, October 4-7, 2017.*, pp. 264–269, Uberlândia-MG.

Leão, J. C.; Brandão, M. A.; Vaz de Melo, P. O. S. & Laender, A. H. F. (2018). Who is really in my social circle? Mining social relationships to improve detection of real communities. *Journal of Internet Services and Applications*. Aceito para publicação.

Liu, X. & Murata, T. (2010). Advanced modularity-specialized label propagation algorithm for detecting communities in networks. *Physica A: Statistical Mechanics and its Applications*, 389(7):1493–1500.

Marsden, P. V. & Campbell, K. E. (1984). Measuring Tie Strength. *Social Forces*, 63(2):482–501.

Moreno, J. (1953). *Who Shall Survive?: Foundations of Sociometry, Group Psychotherapy and Sociodrama*. Nervous and Mental Disease Monograph Series. Beacon House.

Mucha, P. J.; Richardson, T.; Macon, K.; Porter, M. A. & Onnela, J.-P. (2010). Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. *Science*, 328(5980):876–878.

Newman, M. E. (2006a). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

Newman, M. E. J. (2004). Detecting community structure in networks. *The European Physical Journal B*, 38(2):321–330.

Newman, M. E. J. (2006b). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104.

Newman, M. E. J. & Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):26113.

Nicosia, V.; Tang, J.; Mascolo, C.; Musolesi, M.; Russo, G. & Latora, V. (2013). Graph metrics for temporal networks. In *Temporal Networks*, pp. 15–40. Springer, Berlin, Heidelberg.

Nunes, I. O.; Celes, C.; Silva, M.; Vaz de Melo, P. O. S. & Loureiro, A. A. F. (2017). GRM: Group Regularity Mobility Model. In *Proceedings of the 20th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, Miami Beach, USA.

Orke, R. G.; Maillard, P.; Schumm, A.; Staudt, C.; Wagner, D.; Görke, R.; Maillard, P.; Schumm, A.; Staudt, C. & Wagner, D. (2013). Dynamic graph clustering combining modularity and smoothness. *Journal of Experimental Algorithmics*, 18(1):1–5.

Ouyang, B.; Jiang, L. & Teng, Z. (2016). A noise-filtering method for link prediction in complex networks. *PLOS ONE*, 11(1):1–12.

Palla, G.; Barabási, A.-L. & Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446(7136):664–667.

Palla, G.; Derenyi, I.; Farkas, I. & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818.

Papadimitriou, P.; Dasdan, A. & Garcia-Molina, H. (2010). Web graph similarity for anomaly detection. *Journal of Internet Services and Applications*, 1(1):19–30.

Peel, L.; Larremore, D. B. & Clauset, A. (2017). The ground truth about metadata and community detection in networks. *Science Advances*, 3(5).

Peixoto, T. P. & Rosvall, M. (2017). Modelling sequences and temporal networks with dynamic community structures. *Nat. Commun.*, 8(1):582.

Pollner, P.; Palla, G. & Vicsek, T. (2012). Parallel Clustering with CFinder. *Parallel Processing Letters*, 22(01):1240001.

Pons, P. & Latapy, M. (2005). *Computing Communities in Large Networks Using Random Walks*, pp. 284–293. Springer Berlin Heidelberg, Berlin, Heidelberg.

Prat-Pérez, A.; Dominguez-Sal, D.; Brunat, J. M. & Larriba-Pey, J.-L. (2012). Shaping communities out of triangles. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pp. 1677–1681, New York, NY, USA. ACM.

Radicchi, F.; Castellano, C.; Cecconi, F.; Loreto, V.; Parisi, D. & Fisica, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2658–2663.

Raghavan, U. N.; Albert, R. & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, 76(3):1–12.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Rocha, L. E. C.; Masuda, N. & Holme, P. (2017). Sampling of temporal networks: Methods and biases. *Phys. Rev. E*, 96:052302.

Rossetti, G. & Cazabet, R. (2017). Community discovery in dynamic networks: a survey. *CoRR*, abs/1707.03186.

Rosvall, M. & Bergstrom, C. T. (2011). Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PLOS ONE*, 6(4):1–10.

Sah, P.; Singh, L. O.; Clauset, A. & Bansal, S. (2014). Exploring community structure in biological networks with random graphs. *BMC Bioinformatics*, 15(1):220.

Schuetz, P. & Caflisch, A. (2008). Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement. *Physical Review E*, 77(4):046112.

Serrano, M. Á.; Boguná, M. & Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16):6483–6488.

Shi, X.; Adamic, L. A. & Strauss, M. J. (2007). Networks of strong ties. *Physica A: Statistical Mechanics and its Applications*, 378(1):33 – 47.

Spitz, A.; Gimmler, A.; Stoeck, T.; Zweig, K. A. & Horvát, E.-Á. (2016). Assessing low-intensity relationships in complex networks. *PloS one*, 11(4):e0152536.

Šubelj, L. & Bajec, M. (2011). Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction. *Physical Review E*, 83(3):036103.

Teixeira, J.; Robles, G. & González-Barahona, J. M. (2015). Lessons learned from applying social network analysis on an industrial free/libre/open source software ecosystem. *Journal of Internet Services and Applications*, 6(1):14.

Treurniet, J. (2014). A taxonomy and survey of microscopic mobility models from the mobile networking domain. *ACM Comput. Surv.*, 47(1):14:1–14:32.

Vaz de Melo, P. O. S.; Viana, A. C.; Fiore, M.; Jaffrès-Runser, K.; Mouël, F. L.; Loureiro, A. A. F.; Addepalli, L. & Guangshuo, C. (2015). RECAST: Telling Apart Social and Random Relationships in Dynamic Networks. *Performance Evaluation*, 87:19–36.

Wang, L. & Hopcroft, J. (2010). Community structure in large complex networks. In *Proceedings of the 7th International Conference on Theory and Applications of Models of Computation*, pp. 455–466, Berlin, Heidelberg. Springer Berlin Heidelberg.

Wang, M.; Wang, C.; Yu, J. X. & Zhang, J. (2015). Community Detection in Social Networks: An In-depth Benchmarking Study with a Procedure-Oriented Framework. *Proceedings of the VLDB Endowment*, 8(10):998–1009.

Wasserman, S. & Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8 of *Structural Analysis in the Social Sciences*. Cambridge University Press, Cambridge, UK.

Wen, H.; Leicht, E. A. & D'Souza, R. M. (2011). Improving community detection in networks by targeted node removal. *Phys. Rev. E*, 83:1–8.

Xie, J.; Kelley, S. & Szymanski, B. K. (2013). Overlapping Community Detection in Networks : The State-of-the-Art and Comparative Study. *ACM Computing Surveys*, 45(4):43.

Yang, J. & Leskovec, J. (2015). Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213.

Yang, Z.; Algesheimer, R. & Tessone, C. J. (2016). A Comparative Analysis of Community Detection Algorithms on Artificial Networks. *Nature Publishing Group*, (August):1–16.

Yu, P. S.; Han, J. & Faloutsos, C. (2010). *Link Mining: Models, Algorithms, and Applications*. Springer-Verlag New York, New York, NY, USA, 1st edição.

Zaki, M. J. & Wagner Meira, J. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, Cambridge, UK.