MINISTÉRIO DA DEFESA
EXÉRCITO BRASILEIRO
DEPARTAMENTO DE CIÊNCIA E TECNOLOGIA
INSTITUTO MILITAR DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS E COMPUTAÇÃO

MARIANA MAGALHÃES DE MATTOS COELHO

PREDIÇÃO DE *LINKS*: UMA ANÁLISE COMPARATIVA DE MÉTRICAS TOPOLÓGICAS EM REDES DE COAUTORIA

RIO DE JANEIRO
2021

MARIANA MAGALHÃES DE MATTOS COELHO

PREDIÇÃO DE *LINKS*: UMA ANÁLISE COMPARATIVA DE MÉTRICAS TOPOLÓGICAS EM REDES DE COAUTORIA

Dissertação apresentada ao Programa de Pós-graduação em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de Mestre em Ciências em Sistemas e Computação.

Orientador: Claudia Marcela Justel, D.Sc.

Rio de Janeiro

2021

©2021
INSTITUTO MILITAR DE ENGENHARIA
Praça General Tibúrcio, 80 – Praia Vermelha
Rio de Janeiro – RJ CEP: 22290-270

Este exemplar é de propriedade do Instituto Militar de Engenharia, que poderá incluí-lo em base de dados, armazenar em computador, microfilmar ou adotar qualquer forma de arquivamento.

É permitida a menção, reprodução parcial ou integral e a transmissão entre bibliotecas deste trabalho, sem modificação de seu texto, em qualquer meio que esteja ou venha a ser fixado, para pesquisa acadêmica, comentários e citações, desde que sem finalidade comercial e que seja feita a referência bibliográfica completa.

Os conceitos expressos neste trabalho são de responsabilidade do(s) autor(es) e do(s) orientador(es).

Coelho, Mariana Magalhães de Mattos.
Predição de *links*: uma análise comparativa de métricas topológicas em redes de coautoria / Mariana Magalhães de Mattos Coelho. – Rio de Janeiro, 2021.
70 f.

Orientador: Claudia Marcela Justel.

Dissertação (mestrado) – Instituto Militar de Engenharia, Sistemas e Computação, 2021.

1. aplicações de grafos. 2. redes sociais. 3. predição de links. i. Justel, Claudia Marcela (orient.) ii. Título

MARIANA MAGALHÃES DE MATTOS COELHO

Predição de *links*: uma análise comparativa de métricas topológicas em redes de coautoria

Dissertação apresentada ao Programa de Pós-graduação em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de Mestre em Ciências em Sistemas e Computação.

Orientador: Claudia Marcela Justel.

Aprovado em Rio de Janeiro, 26 de novembro de 2021, pela seguinte banca examinadora:

Prof. **Ronaldo Ribeiro Goldschmidt** - D.Sc. do IME - Presidente

Prof. **Claudia Marcela Justel** - D.Sc. do IME

Prof. **Carla Silva Oliveira** - D.Sc. da ENCE - IBGE

Rio de Janeiro
2021

Ao Instituto Militar de Engenharia, alicerce da minha formação e aperfeiçoamento , ao Serviço Geológico do Brasil, pela oportunidade de crescimento pessoal e profissional e, especialmente, à minha querida avó, Maria Orchidea Magalhães Trajano, minha maior incentivadora (in memoriam).

AGRADECIMENTOS

Agradeço a Deus e ao meu anjo da guarda por estarem presentes em todos os momentos da minha vida, à minha querida família, em especial aos meus pais, Maria Eldair e Oscar, pelo apoio e amor incondicionais e à Maica, minha segunda mãe, por ter me criado com muito amor e carinho.

Agradeço ainda aos professores que me incentivaram, em especial à minha orientadora, Claudia Marcela Justel, por sua atenção, disponibilidade e exemplo. Agradeço também ao colega Argus, por toda ajuda nos momentos necessários e a cada um dos companheiros deste curso, que trilharam esta jornada comigo.

Por fim, agradeço ao Instituto Militar de Engenharia pelo apoio proporcionado, principalmente pela máquina disponibilizada pelo Laboratório de Alto Desempenho da Defesa Cibernética para que os experimentos pudessem ser realizados, e ao Serviço Geológico do Brasil pela oportunidade de capacitação.

"Lembre-se de que as pessoas podem tirar tudo de você,
menos o seu conhecimento."
(Albert Einstein)

RESUMO

O problema denominado predição de *links* consiste em estimar a existência do surgimento de arestas entre nós de um grafo que representa uma rede. Diversas abordagens para resolver esse problema foram propostas nos últimos anos. Dentre as diversas abordagens do problema, neste trabalho consideramos a topológica. Para resolver essa abordagem do problema, foram propostas diferentes métricas, por exemplo as métricas topológicas locais em duas versões: tradicional e aos pares. Até o nosso conhecimento, não foi realizada uma comparação do desempenho de ambas versões das métricas topológicas locais tradicional e aos pares. O objetivo deste trabalho é comparar métricas topológicas locais que denominaremos 'tradicional' e 'aos pares' propostas na literatura através de experimentos em redes de coautoria. Apresentamos resultados dos experimentos realizados, utilizando quatro métricas topológicas locais em duas versões, em seis redes reais, assim como as conclusões obtidas a respeito da comparação.

Palavras-chave: aplicações de grafos. redes sociais. predição de links.

ABSTRACT

The problem called link prediction consists of estimating the existence of the appearance of edges between nodes of a graph that represents a network. Several approaches to solving this problem has been proposed in recent years. Among those approaches, in this work we consider the topological one. In order of solve this version of the problem, different metrics were proposed and among them, the topological local metrics denoted as tradicional and pairwise were proposed. To our knowledge, there are no researches about comparing performances of topological local metrics in these two versions. The objective of this work is to compare topological metrics that we will call 'traditional' and 'pairwise' proposed in the literature through experiments in co-authorship networks. We present results of the experiments in six real networks and compare the performance of the solution of the link prediction problem by using four local topological metrics in both versions, as well as our conclusions about it.

Keywords: graph applications. social networks. link prediction.

LISTA DE ILUSTRAÇÕES

Figura 1 – Detecção de comunidades. 17
Figura 2 – Medidas de centralidade. 18
Figura 3 – Predição de *links*. 18
Figura 4 – Grafo $G = (V, E), |V| = 5, |E| = 4$. 22
Figura 5 – Exemplo de grafo bipartido. 23
Figura 6 – Adaptado da Figura 1. 25
Figura 7 – Adaptado da Figura 1. 25
Figura 8 – Projeções de um grafo bipartido. 36
Figura 9 – O primeiro grafo representa G_t, as ligações existentes até o instante t, e o segundo $G_{t'}$, as ligações existentes até o instante t', para $t < t'$. Esta figura representa a predição de *links* tradicional. Adaptado da Figura 1. 37
Figura 10 – O primeiro grafo representa G_t, as ligações existentes até o instante t, e o segundo $G_{t'}$, as ligações existentes até o instante t', para $t < t'$. Esta figura representa a predição de *links* aos pares. Adaptado da Figura 1. 37
Figura 11 – G_{2011}. 43
Figura 12 – G_{2014}. 43

LISTA DE TABELAS

Tabela 1 – Tabela comparativa de trabalhos relacionados. 37
Tabela 2 – Matriz de Confusão de um Classificador - problema com 2 classes. . . . 39
Tabela 3 – Métricas utilizadas nos experimentos para a versão tradicional. 41
Tabela 4 – Métricas utilizadas nos experimentos para a versão aos pares 'ou'. . . . 41
Tabela 5 – Métricas utilizadas nos experimentos para a versão aos pares 'e'. 42
Tabela 6 – Informações das redes de coautoria obtidas da Plataforma Lattes - CNPq. V_{2011} e E_{2011} são os conjuntos de autores e publicações até 2011. E_{2014} é o conjunto de publicações até 2014 dos autores em V_{2011}. 45
Tabela 7 – Informações das redes de coautoria do *ArXiv*. V_{1997} e E_{1997} são os conjuntos de autores e publicações até 1997. E_{1998} é o conjunto de publicações até 2018 dos autores em V_{1997}. 45
Tabela 8 – Resultados para G_{2011}, G_{2014} na rede Lattes do Experimento 1.1 47
Tabela 9 – Resultados para G_{2011}, G_{2014} na rede Lattes do Experimento 1.2 48
Tabela 10 – Resultados para G_{1997}, G_{1998} no *dataset* gr-qc do Experimento 2.1 . . . 50
Tabela 11 – Resultados para G_{1997}, G_{1998} no *dataset* cond-mat do Experimento 2.1 51
Tabela 12 – Resultados para G_{1997}, G_{1998} no *dataset* astro-ph do Experimento 2.1 . 52
Tabela 13 – Resultados para G_{1997}, G_{1998} no *dataset* hep-ph do Experimento 2.1 . . 53
Tabela 14 – Resultados para G_{1997}, G_{1998} no *dataset* hep-th do Experimento 2.1 . . 54
Tabela 15 – Resultados para G_{1997}, G_{1998} no *dataset* gr-qc do Experimento 2.2 . . . 55
Tabela 16 – Resultados para G_{1997}, G_{1998} no *dataset* cond-mat do Experimento 2.2 57
Tabela 17 – Resultados para G_{1997}, G_{1998} no *dataset* astro-ph do Experimento 2.2 . 58
Tabela 18 – Resultados para G_{1997}, G_{1998} no *dataset* hep-ph do Experimento 2.2 . . 59
Tabela 19 – Resultados para G_{1997}, G_{1998} no *dataset* hep-th do Experimento 2.2 . . 60
Tabela 20 – Tabela comparativa das versões aos pares 'ou' e 'e' da rede Lattes . . . 62
Tabela 21 – Tabela comparativa das versões aos pares 'ou' e 'e' das redes do *ArXiv*. 63

LISTA DE ABREVIATURAS E SIGLAS

VC	Vizinhos Comuns
JS	Similaridade de Jaccard
LP	Ligação Preferencial
AA	Adamic-Adar
VC*	Vizinhos Comuns aos pares utilizando a métrica ‘ou’
JS*	Similaridade de Jaccard aos pares utilizando a métrica ‘ou’
LP*	Ligação Preferencial aos pares utilizando a métrica ‘ou’
AA*	Adamic-Adar aos pares utilizando a métrica ‘ou’
VC**	Vizinhos Comuns aos pares utilizando a métrica ‘e’
JS**	Similaridade de Jaccard aos pares utilizando a métrica ‘e’
LP**	Ligação Preferencial aos pares utilizando a métrica ‘e’
AA**	Adamic-Adar aos pares utilizando a métrica ‘e’
VCP	Perfil de colocação de vértices
MRLP	Multi-relational link prediction
VP	Verdadeiro positivo
FP	Falso positivo
FN	Falso negativo
VN	Verdadeiro negativo

LISTA DE SÍMBOLOS

α	Letra grega Alfa
β	Letra grega Beta
γ	Letra grega Gama minúscula
Γ	Letra grega Gama maiúscula
λ	Letra grega Lambda
$\in$	Pertence
$\bigcup$	União
$\bigcap$	Interseção
$/$	Tal que
$\sum$	Somatório
∞	Infinito
$\times$	Produto cartesiano
$\emptyset$	Conjunto vazio

SUMÁRIO

1 INTRODUÇÃO 15
1.1 CONTEXTUALIZAÇÃO E MOTIVAÇÃO 16
1.2 PROBLEMA DE PESQUISA 19
1.2.1 LIMITAÇÃO DAS MÉTRICAS ANALISADAS 20
1.3 OBJETIVO 20
1.4 METODOLOGIA 21
1.5 CONTRIBUIÇÕES ESPERADAS 21
1.6 ORGANIZAÇÃO DO TEXTO 21

2 FUNDAMENTAÇÃO TEÓRICA 22
2.1 CONSIDERAÇÕES INICIAIS 22
2.2 CONCEITOS EM TEORIA DE GRAFOS 22
2.3 PREDIÇÃO DE *LINKS* 24
2.3.1 ABORDAGENS 25
2.3.2 MÉTODOS TOPOLÓGICOS BASEADOS EM SIMILARIDADE 26
2.3.2.1 ABORDAGENS LOCAIS 27
2.3.2.2 ABORDAGENS GLOBAIS 29
2.3.3 ESTRUTURA DE ORDEM MAIS COMPLEXA 30
2.3.4 MÉTODOS PARA PREDIÇÃO DE LINKS AOS PARES 31

3 TRABALHOS RELACIONADOS 34
3.1 PREDIÇÃO DE *LINKS* 34
3.2 REDES HETEROGÊNEAS E MÉTRICAS TOPOLÓGICAS 35
3.3 CONCLUSÕES 37

4 METODOLOGIA 38
4.1 ABORDAGEM PROPOSTA 38
4.1.1 MEDIDAS DE QUALIDADE 39
4.1.2 PREDITOR RANDÔMICO 40
4.2 DESCRIÇÃO DAS MÉTRICAS 40
4.3 *DATASETS* 42
4.3.1 DEFINIÇÃO DOS *DATASETS* 44
4.4 AMBIENTE DE APOIO À EXPERIMENTAÇÃO 45

5 EXPERIMENTOS E RESULTADOS 46
5.1 CONSIDERAÇÕES INICIAIS 46
5.2 EXPERIMENTO 1 46

5.2.1 DESCRIÇÃO DO EXPERIMENTO 1 46
5.2.2 RESULTADOS OBTIDOS DO EXPERIMENTO 1 47
5.3 EXPERIMENTO 2 48
5.3.1 DESCRIÇÃO DO EXPERIMENTO 2 48
5.3.2 RESULTADOS OBTIDOS DO EXPERIMENTO 2 49

6 JUSTIFICATIVA E COMPARAÇÃO 62

7 CONCLUSÃO 65

REFERÊNCIAS 67

1 INTRODUÇÃO

Muitos sistemas sociais, biológicos e de informação podem ser descritos por redes cujos nós representam indivíduos, elementos biológicos (proteínas, genes, etc.), computadores, usuários da web e assim por diante, e *links* denotam as relações ou interações entre nós (Lü; ZHOU, 2011). Um *link* é uma conexão entre dois vértices em uma rede. Este conceito simples pode ser usado para representar sistemas extremamente complexos por meio da interação entre eles representado por um grande número de elementos formando redes complexas.

O estudo de redes complexas tornou-se um foco comum de muitos ramos da Ciência. Muitos esforços têm sido feitos para entender a evolução das redes, as relações entre topologias e funções e as características da rede (Lü; ZHOU, 2011).

Nos últimos anos, percebeu-se um grande avanço na utilização e popularização das redes sociais. Algumas redes sociais envolvendo um único tipo de nós e *links* podem ser representadas como redes homogêneas, enquanto as redes sociais contendo informações abundantes sobre quem, onde, quando e o que, podem ser denotadas como redes heterogêneas (ZHANG; YU, 2014).

Para analisar redes sociais, existem diversas abordagens, como medidas de centralidade (NEWMAN, 2006a), detecção de comunidades (FORTUNATO, 2010) e a predição de *links* (LIBEN-NOWELL; KLEINBERG, 2007). A predição de *links* é um tipo de análise de redes sociais no caso dinâmico, ou seja, a rede muda ao longo do tempo. O problema de predição de *links* visa inferir o comportamento do processo da formação de *links* de rede, prevendo relacionamentos perdidos ou futuros com base nas conexões atualmente observadas.

A predição de *links* tornou-se uma área de estudo atraente, pois permite prever como as redes evoluirão em um futuro próximo, podendo antecipar relações ainda não identificadas em uma organização. Métodos típicos para predição de *links* usam a topologia da rede para prever o futuro mais provável ou conexões ausentes entre um par de nós (CLAUSET; MOORE; NEWMAN, 2008).

Na biologia, a predição de *links* é usada para identificar novas interações entre genes, doenças e medicamentos dentro de redes de interação (LIN et al., 2018). E em aplicativos relacionados à segurança, pode ser usada para identificar grupos ocultos de terroristas e criminosos.

Entretanto, a evolução da rede é frequentemente mediada por estruturas mais complexas envolvendo mais que dois nós. Por exemplo, subgrafos completos de tamanho 3 (também chamados triângulos) são a chave para a estrutura das redes sociais, mas a

estrutura tradicional de predição de *links* não prevê diretamente essas estruturas (NASSAR; BENSON; GLEICH, 2019a).

Segundo (NASSAR; BENSON; GLEICH, 2019a), há evidências crescentes de que a organização e evolução das redes são centradas em torno de interações mais complexas envolvendo mais de dois nós. No caso de redes sociais, triângulos (subgrafos completos de três nós) são extremamente comuns devido a vários mecanismos sociológicos que impulsionam o fechamento triádico.

Para atender à formação de estruturas mais complexas, (NASSAR; BENSON; GLEICH, 2019a) propuseram uma nova tarefa de predição de *links* chamada predição de *links* aos pares que tem como objetivo prever o surgimento de novos triângulos, nos quais se tem o propósito de encontrar quais nós são mais propensos a formar um triângulo com uma determinada aresta.

Como exemplos de plataformas de aplicações de predição de *links*, têm-se a Amazon, o LinkedIn, a Netflix e o Facebook. Nas redes sociais online de amizades, como, por exemplo, o Facebook, a predição de que duas pessoas irão formar uma conexão pode ser usada para recomendação de amizade (BACKSTROM; LESKOVEC, 2011).

De forma similar, a predição de novos *links* entre usuários e itens em plataformas comerciais, como Amazon e Netflix, pode ser usada para recomendação de produtos (GOMEZ-URIBE; HUNT, 2015). A Netflix poderia sugerir um filme para um cliente e a Amazon poderia sugerir um livro para um usuário. Já o LinkedIn poderia sugerir uma vaga de emprego para um usuário cadastrado em seu banco de dados.

Existem vários cenários em que o problema de predição de *links* aos pares é natural, como recomendar um novo amigo para um casal em uma rede social online, recomendar um filme para um casal em um site de vídeo ou fazer predição de um medicamento eficaz dado um par doença-gene.

1.1 Contextualização e Motivação

Uma rede social é representada por um grafo denso e cada vértice do grafo representa um indivíduo ou organização e uma aresta representa algum tipo de interação entre os nós, como, por exemplo, uma relação de amizade ou uma colaboração entre eles. A rede social cresce e muda rapidamente ao longo do tempo através do surgimento de novas arestas que representam novas associações na estrutura social e de novos nós que configuram novos elementos (LIBEN-NOWELL; KLEINBERG, 2003).

O objetivo da detecção de comunidades em grafos (FORTUNATO, 2010) é identificar conjuntos de nós que estão fortemente conectados entre si, mas fracamente conectados aos outros elementos no grafo (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

A Figura 1 representa a rede de Zachary, uma rede de um clube de karatê dos EUA. Os nós representam os integrantes do clube e as arestas representam as relações de amizade entre eles. Houve um conflito entre o instrutor e o presidente do clube e alguns membros do clube apoiaram o instrutor e outros defenderam o presidente. Existem 34 indivíduos na rede. O nó 1 representa o instrutor e o nó 34 representa o presidente.

Esta rede de pequeno porte foi usada para testar algoritmos de detecção de comunidades. Observa-se na Figura 1 que o resultado das 4 comunidades representadas por cores diferentes dos nós que a compõem foram obtidas por um algoritmo de detecção que posiciona os nós 1 e 34 em comunidades diferentes.

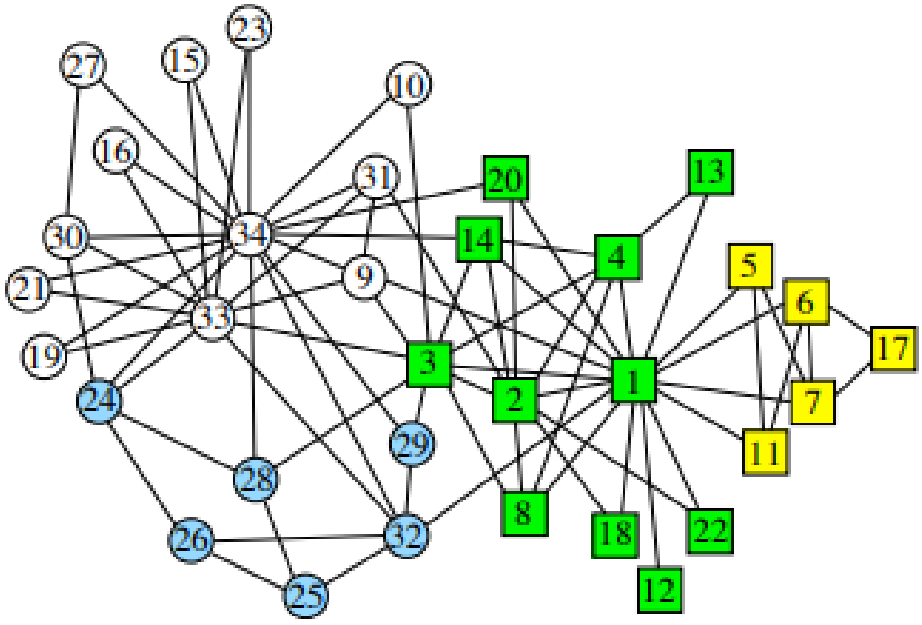

Figura 1 – Detecção de comunidades (FORTUNATO, 2010).

Existem distintas medidas de centralidade que permitem identificar quais os nós mais importantes em uma rede, como por exemplo: grau, proximidade, intermediação, hubs e autoridades. A Figura 2 representa uma rede de colaborações entre cientistas, com 379 indivíduos (NEWMAN, 2006a). É uma rede de coautoria de artigos científicos. Os *links* representam trabalhos publicados em coautoria entre pares de cientistas. Os nós que estão em vermelho são os nós centrais da rede. Os diâmetros dos vértices indicam a centralidade da comunidade. Os vértices com as centralidades mais altas são destacados. A centralidade da comunidade destaca os vértices que são centrais em suas comunidades locais.

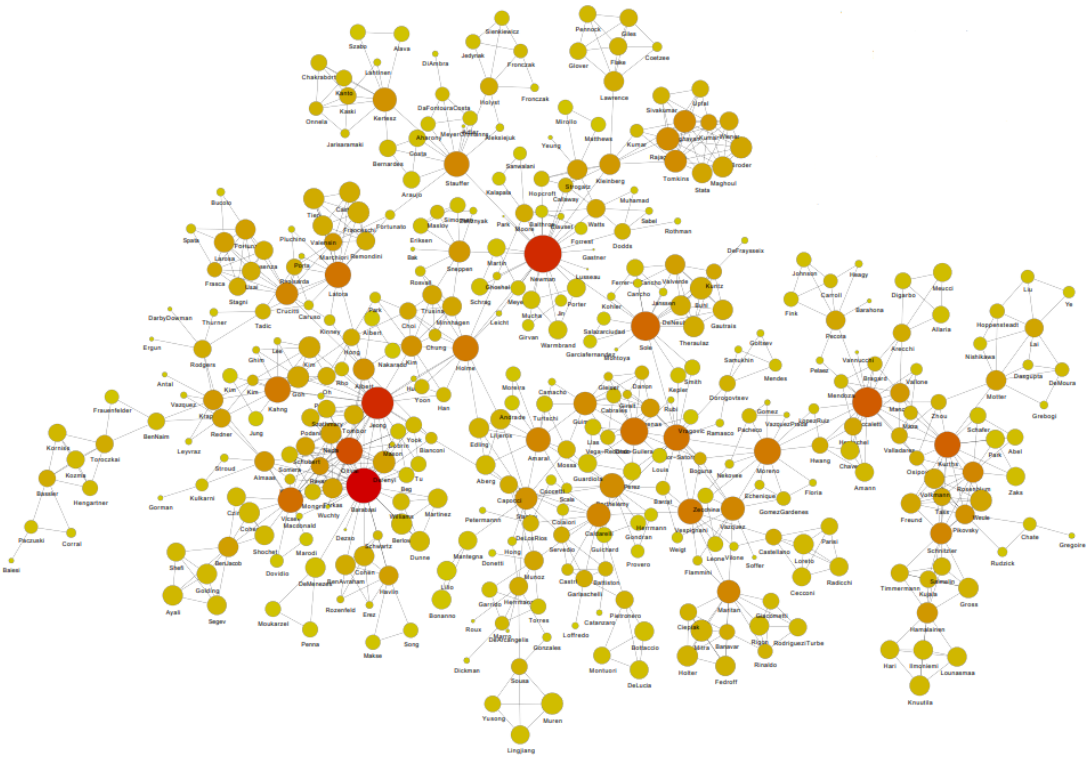

Figura 2 – Medidas de centralidade (NEWMAN, 2006b)

A Figura 3 (FLORENTINO; GOLDSCHMIDT, 2017) representa a predição de *links* tradicional em uma rede contendo informações correspondentes aos anos 2000, 2001 e 2002. Ou seja, dada a rede composta pelas informações dos anos 2000 e 2001, será possível prever as conexões que ocorreriam nesta rede no ano de 2002?

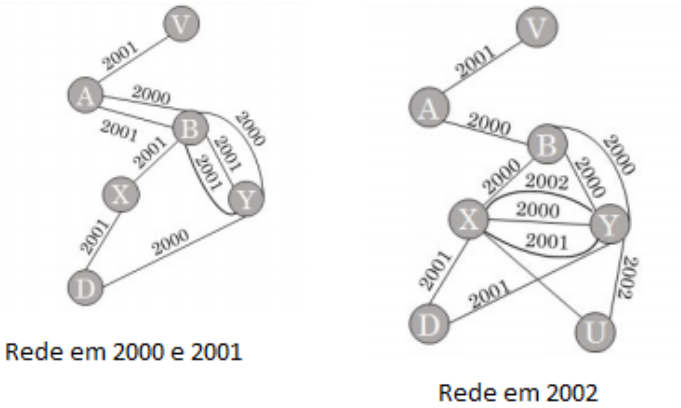

Figura 3 – Predição de *links* (FLORENTINO; GOLDSCHMIDT, 2017)

Os tipos de informações geralmente utilizadas para resolver o problema de predição de *links* são: topológica (LIBEN-NOWELL; KLEINBERG, 2007), contextual (VALVERDE-REBAZA; LOPES, 2013), temporal (VALVERDE-REBAZA; LOPES, 2013) e combinação das anteriores (VALVERDE-REBAZA; LOPES, 2013).

As informações topológicas são aquelas que descrevem a estrutura do grafo que representa a rede, sem considerar a semântica inerente ao contexto da aplicação da rede social. Por exemplo, o grau de um nó ou a quantidade de vizinhos comuns a um par de vértices são informações de natureza tipicamente topológica. Este tipo de dado não está explicitamente disponível na rede: deve ser calculado a partir da estrutura da rede (MUNIZ; GOLDSCHMIDT; CHOREN, 2018).

Já as informações contextuais são aquelas que expressam características dos elementos das redes e de suas interações. Este tipo de informação contempla a semântica que está no âmbito da aplicação da rede social. A existência de tais informações é diretamente dependente da rede analisada e o conjunto de dados contextuais disponíveis varia de uma rede para outra por ser dependente do contexto (MUNIZ; GOLDSCHMIDT; CHOREN, 2018).

As informações temporais são aquelas que verificam em que momento ocorreram as conexões entre os elementos da rede, ou seja, quando ocorreram as interações sociais. As informações temporais verificam os dados cronológicos importantes relacionados a aspectos topológicos e contextuais da rede social (MUNIZ; GOLDSCHMIDT; CHOREN, 2018).

E a combinação das anteriores considera de forma conjunta dois ou mais tipos de informações: topológicas, contextuais e temporais; ou seja, combinando propriedades obtidas da estrutura do grafo, as informações adicionais da rede e o momento em que ocorreram as conexões sociais (MUNIZ; GOLDSCHMIDT; CHOREN, 2018).

1.2 Problema de Pesquisa

Dentre as diferentes propostas que existem na literatura para resolver o problema de predição de *links* usando informações topológicas, destacam-se duas versões: tradicional (LIBEN-NOWELL; KLEINBERG, 2007) e aos pares (NASSAR; BENSON; GLEICH, 2019a).

O problema da predição de *links* tradicional (LIBEN-NOWELL; KLEINBERG, 2007) consiste em identificar pares de nós que formarão uma ligação no futuro, enquanto a rede evolui no tempo. O problema pode ser descrito pela seguinte pergunta: dada uma rede no tempo t, é possível inferir quais novas conexões entre seus membros (não conectados) são prováveis de ocorrer no futuro próximo t' $(t < t')$?

O problema da predição de *links* aos pares (NASSAR; BENSON; GLEICH, 2019a)

consiste em identificar um nó que formará uma ligação no futuro com os nós de uma aresta, formando um triângulo, enquanto a rede evolui no tempo. O problema pode ser descrito pela seguinte pergunta: dada uma aresta (u, v) na rede no tempo t, quais nós são prováveis de se conectar aos nós da aresta (u, v) no futuro próximo t' $(t < t')$? É importante ressaltar que os nós prováveis de se conectar aos nós da aresta não podem estar conectados simultaneamente à aresta, mas podem estar conectados a apenas um nó da aresta ou a nenhum nó da aresta (u, v) em questão.

O foco principal deste trabalho é a abordagem não supervisionada baseada em informações topológicas do problema de predição de *links*. Segundo a literatura, as métricas topológicas mais usadas, que são chamadas neste trabalho na versão 'tradicional', foram propostas por (LIBEN-NOWELL; KLEINBERG, 2007). As que são chamadas nesta dissertação de versão 'aos pares' foram propostas recentemente (NASSAR; BENSON; GLEICH, 2019a) e (NASSAR; BENSON; GLEICH, 2020) e determinam um novo método para resolver o problema. Nos trabalhos de Nassar et al., as versões aos pares das métricas topológicas foram usadas para melhorar a métrica topológica global conhecida na literatura como PageRank (BRIN; PAGE, 1998).

1.2.1 Limitação das métricas analisadas

Pela pesquisa realizada, não foi encontrada uma análise comparativa das métricas locais tradicionais e aos pares. Provavelmente, porque a versão aos pares é bem recente e foi usada no contexto de melhorar o desempenho do PageRank, que é uma métrica global.

1.3 Objetivo

A presente dissertação levanta a hipótese de que as métricas topológicas aos pares podem ser mais vantajosas comparadas às métricas tradicionais na tarefa de predição de *links*. O objetivo geral deste trabalho é determinar vantagens e desvantagens da utilização das métricas topológicas tradicionais e aos pares fazendo uma análise comparativa em redes homogêneas.

Para cumprir o objetivo geral enunciado acima, este trabalho possui ainda os seguintes objetivos específicos:

- Formular um método para demonstração da hipótese levantada.
- Realizar experimentos em redes reais de coautoria.
- Fazer uma análise comparativa dos resultados obtidos ao utilizar as métricas topológicas tradicionais e aos pares em redes de coautoria.

1.4 Metodologia

De forma a atender os objetivos desta dissertação, foi aplicado o método de predição de *links* que resgata apenas informações topológicas da rede. Seis redes de coautoria serão utilizadas para realizar os experimentos. Uma rede de menor tamanho obtida a partir de informações da Plataforma Lattes do CNPq e cinco redes muito utilizadas para testar métodos de predição de *links*, obtidas a partir do repositório *ArXiv*. A partir deste resgate, são calculados todos os pares de nós ainda não conectados na rede para as métricas tradicionais e todos os pares formados por uma aresta e um nó, sendo que o nó não seja adjacente simultaneamente às duas extremidades da aresta, para a métrica aos pares. O método proposto retornará um ranqueamento dos pares (nó, nó), (nó, aresta) segundo a versão tradicional e aos pares, em ordem decrescente.

A seguir, é realizada a análise dos resultados obtidos por cada versão de quatro métricas locais (Vizinhos Comuns, Similaridade de Jaccard, Ligação Preferencial e Adamic-Adar), nas versões tradicional e aos pares para cada rede analisada. Para tal fim, será utilizada a matriz de confusão e o cálculo das medidas de qualidade que permitem comparar os resultados obtidos com as duas versões das métricas, tradicional e aos pares. As redes de coautoria utilizadas foram as seguintes: uma rede Lattes e cinco redes do *ArXiv* (astro-ph, cond-mat, gr-qc, hep-ph e hep-th).

1.5 Contribuições Esperadas

A contribuição esperada como resultado deste trabalho é fazer uma comparação do desempenho das métricas topológicas locais nas versões 'tradicional' e 'aos pares' em redes de coautoria, determinando vantagens e desvantagens de cada tipo de métrica.

1.6 Organização do Texto

A presente pesquisa está organizada em sete capítulos. Este capítulo apresenta a introdução do trabalho, com a exposição da motivação, a caracterização do problema, objetivo, metodologia e contribuições esperadas. No Capítulo 2 será apresentada a fundamentação teórica oportuna. Os trabalhos relacionados ao assunto tratado nesta pesquisa serão descritos no Capítulo 3. Já no Capítulo 4 será tratada a metodologia para comparação das métricas tradicional e aos pares. No Capítulo 5 serão apresentados os experimentos realizados e os resultados obtidos. No Capítulo 6 serão relatadas a justificativa e a comparação das métricas. Por fim, no Capítulo 7, será apresentada a conclusão do trabalho, destacando as contribuições e as sugestões de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Considerações iniciais

Nas próximas subseções serão apresentados os conceitos básicos que foram utilizados na pesquisa. Serão descritos os conceitos de teoria dos grafos, a tarefa principal desta dissertação que é a predição de *links* tradicional e a predição de *links* aos pares e as abordagens utilizadas para o tratamento do problema da predição de *links*.

2.2 Conceitos em teoria de grafos

Um **grafo** G é um par ordenado $G = (V, E)$, em que V é um conjunto de vértices ou nós e E é um conjunto de arestas entre pares de elementos do conjunto V. Uma aresta entre dois nós x e y é denotada por (x, y). O número de nós no grafo, também conhecido como tamanho do grafo, é denotado como $|V|$. O número de arestas é denotado como $|E|$ (TRUDEAU, 1993).

Em **grafos não direcionados**, o grau de um nó é definido como o número de arestas conectadas ao nó e será indicado como $\mid \Gamma(x) \mid$. A Figura 4 representa um grafo não direcionado com cinco vértices e quatro arestas.

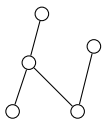

Figura 4 – Grafo $G = (V, E), |V| = 5, |E| = 4$.

O **grau médio** de um grafo, denotado como $\overline{d}$, é igual à média dos graus de todos os seus nós. Um **laço** é uma aresta conectando um nó a si próprio (SZWARCFITER, 2018). Um **grafo simples** é definido como um grafo sem laço e com não mais de uma aresta ou arco entre cada par de vértices.

Um **caminho** em um grafo é uma sequência finita ou infinita de nós conectados por arestas (EASLEY; KLEINBERG, 2010). O caminho é denotado como: v_1 ... v_k tal que $(v_i, v_{i+1}) \in E$, $1 \leq i \leq k-1$. O **comprimento do caminho** é o número de arestas no caminho (SZWARCFITER, 2018). O **caminho mais curto** entre dois vértices é o caminho com o menor comprimento entre esses vértices. Vários caminhos mais curtos para um par de vértices podem existir.

Um grafo $G = (V, E)$ é chamado **conexo** se existe um caminho entre cada par de nós $x, y \in V$ (SZWARCFITER, 2018). Se o grafo não for conexo, ele é composto de **componentes conexas**. Uma componente conexa é um subgrafo conexo maximal de G. Um grafo conexo possui apenas uma componente conexa. Se uma das componentes tiver um número significativamente maior de nós em comparação com as outras componentes conexas, geralmente é chamado de **componente principal ou gigante**.

Normalmente, usam-se letras minúsculas, como x, y, z, para indicar um nó em um grafo e as arestas são representadas pela letra e. $\Gamma(u)$ é o conjunto de vizinhos de $u \in V$.

- $\Gamma(u) = \{v \in V \ / \ (u, v) \in E \}$, vizinhos do vértice u.

Dado um grafo $G(V, E)$, a **matriz de adjacências** A = $(A_{i,j})$ é uma matriz n x n tal que:

- $A_{i,j} = 1$, se $(x,y) \in E$
- $A_{i,j} = 0$, caso contrário;

ou seja, $A_{i,j} = 1$ quando os vértices v_i,v_j forem adjacentes e $A_{i,j} = 0$, caso contrário (SZWARCFITER, 2018).

Um grafo é **bipartido** se seu conjunto de vértices puder ser particionado em dois conjuntos disjuntos não vazios de maneira tal que cada aresta do grafo possui extremidades em cada um dos conjuntos da partição (SZWARCFITER, 2018). A Figura 5 representa um grafo bipartido.

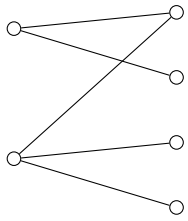

Figura 5 – Exemplo de grafo bipartido.

Um grafo G com n nós é **denso** se o número de arestas é próximo ao seu número máximo ($\frac{n(n-1)}{2}$). Já um grafo **esparso** é um grafo com poucas arestas.

Um subgrafo é dito **induzido** de G se ele tem todas as arestas que estão em G considerando o mesmo conjunto de vértices.

Um grafo **completo** é um grafo que possui todas as arestas possíveis. Denotamos por K_n o grafo completo com n nós. P_n é um caminho com n vértices e tríades significam o subconjunto de três vértices em G.

Dado um grafo $G = (V, E)$, a **transitividade** (*transitivity*) de G, denotada por $T(G)$, é definida por um quociente entre o número de K_3 (triângulos) que existem no grafo e o número de pares de vértices que formam um caminho P_3. Assim, n_3 é o número de triângulos em G, p_3 é o caminho com três vértices. A Equação 2.1 apresenta a fórmula para o cálculo de $T(G)$.

$$T(G) = \frac{3 \; . \; n_3}{p_3} \tag{2.1}$$

Seja $G = (V, E)$, o **coeficiente de agrupamento** (*clustering*) de um vértice $v \in V$, denotado por c_v, determina quão perto os seus vizinhos estão de serem um subgrafo completo em G. A Equação 2.2 mostra como calcular c_v, onde $T(v)$ é o número de triângulos através do nó v e $|\Gamma(v)|$ é o grau de v.

$$c_v = \frac{2T(v)}{|\Gamma(v)|(|\Gamma(v)| - 1)} \tag{2.2}$$

O **coeficiente de agrupamento médio** de $G = (V, E)$ (*average clustering*), denotado por $C(G)$, corresponde, segundo (WATTS; STROGATZ, 1998), ao valor médio dos coeficientes de agrupamento de todos os vértices. A Equação 2.3 mostra como calcular $C(G)$ para o grafo $G = (V, E)$, onde $|V| = n$,

$$C(G) = \frac{1}{n} \sum_{v \in V} c_v \tag{2.3}$$

2.3 Predição de *links*

Redes sociais são objetos altamente dinâmicos, que crescem e mudam rapidamente ao longo do tempo pela adição de novas arestas e vértices, de acordo com o surgimento de novas interações, no grafo original.

O problema de predição de *links* é um problema relacionado com a evolução da rede ao longo do tempo. E pode ser definido da seguinte forma: dado um retrato instantâneo de uma rede num tempo t, o problema de predição de *links* procura prever com certa precisão arestas que serão adicionadas nessa rede durante o intervalo de tempo entre t e um tempo futuro t' (LIBEN-NOWELL; KLEINBERG, 2007).

O problema da predição de *links* aos pares (NASSAR; BENSON; GLEICH, 2019a) é identificar um nó que formará uma ligação no futuro com os nós de uma aresta, formando um triângulo, enquanto a rede evolui no tempo. Descrevendo o problema com uma pergunta seria: dado uma aresta (u, v) na rede no tempo t, quais nós são prováveis de se conectar aos nós da aresta (u, v) no futuro próximo t'?

Para esse conceito, foram dadas duas definições diferentes que denotaremos $\Gamma^*((u,v))$, que se refere à métrica 'ou' (NASSAR; BENSON; GLEICH, 2020) e $\Gamma^{**}((u,v))$, que se refere à métrica 'e' (NASSAR; BENSON; GLEICH, 2019a). Foi definido:

$$\begin{aligned}\Gamma^*((u,v)) &= \{\ z \in V \mid z \text{ é adjacente a } u \text{ ou } z \text{ é adjacente a } v\} \\ &= \Gamma(u) \cup \Gamma(v) \setminus \{u,v\}.\end{aligned}$$

$\Gamma^{**}((u,v)) = \{\ z \in V \mid z \text{ é adjacente a } u \text{ e } z \text{ é adjacente a } v \text{ e } (u,v) \in E\} = \Gamma(u) \cap \Gamma(v)$.

Exemplo 1: A Figura 6 apresenta o grafo G_1, para o qual o vértice b possui vizinhança $\Gamma(b) = \{a, c, d, h\}$.

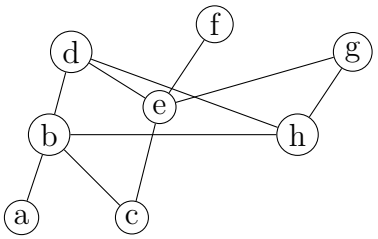

Figura 6 – Adaptado da Figura 1 (NASSAR; BENSON; GLEICH, 2019a).

Exemplo 2: A Figura 7 mostra um grafo G_2 para o qual a aresta (c,e) tem como vizinhança os conjuntos $\Gamma^*(c,e) = \Gamma(c) \cup \Gamma(e) - \{c,e\} = \{b,e\} \cup \{c,d,f,g\} - \{c,e\} = \{b,c,d,e,f,g\} - \{c,e\} = \{b,d,f,g\}$, utilizando a métrica 'ou' e $\Gamma^{**}(c,e) = \Gamma(c) \cap \Gamma(e) = \{b,e\} \cap \{c,d,f,g\} = \emptyset$, utilizando a métrica 'e'.

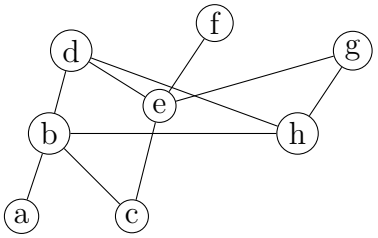

Figura 7 – Adaptado da Figura 1 (NASSAR; BENSON; GLEICH, 2019a).

2.3.1 Abordagens

As abordagens utilizadas para resolução do problema de predição de *links* são as seguintes: supervisionada e não supervisionada. Na abordagem não supervisionada (LIBEN-NOWELL; KLEINBERG, 2007), são calculadas as métricas de similaridade entre pares de vértices não conectados. O resultado dessas métricas são os *scores*, que serão ranqueados em ordem decrescente para verificar os pares de vértices mais prováveis de

se conectarem no futuro próximo, pois quanto maior o valor do *score*, supõe-se que as chances destes nós conectarem-se no futuro são maiores.

Na abordagem não supervisionada, a rede é dividida em período de treino e período de teste. Os pares de vértices não ligados são extraídos da base de dados resultante do período de treino e a ocorrência de uma ligação entre estes é verificada no período de testes.

Já na abordagem supervisionada, o problema de predição de *links* é transformado em uma classificação binária. As técnicas aplicadas para esta estratégia incluem algoritmos de classificação, como, por exemplo, árvores de decisão e bayesiano ingênuo (HASAN et al., 2006).

Ambas as abordagens, supervisionada e não supervisionada, utilizam informações resgatadas da rede para calcular a similaridade entre pares de vértices. O foco deste trabalho é na abordagem não supervisionada.

2.3.2 Métodos Topológicos Baseados em Similaridade

Os métodos baseados em similaridade pressupõem que os nós tendem a formar *links* com outros nós. Esses métodos decorrem da hipótese de que dois nós são semelhantes se forem conectados a nós semelhantes ou estão próximos na rede de acordo com uma determinada função distância. Essas abordagens definem uma função $s(x, y)$ que atribui uma pontuação conhecida como semelhança para cada par de nós x e y (MARTÍNEZ; BERZAL; CUBERO, 2016).

Esta pontuação é calculada para cada par de nós, geralmente aqueles com *links* não observados entre eles. Pares de nós são classificados em ordem decrescente com base em suas pontuações de similaridade, portanto, as arestas no topo da classificação são consideradas mais prováveis de estarem presentes no conjunto de *links* ausentes (MARTÍNEZ; BERZAL; CUBERO, 2016).

A definição de similaridade não é uma tarefa trivial, pois possui um componente heurístico. A função de similaridade pode variar entre redes, mesmo sendo do mesmo domínio de aplicação. Como um resultado não surpreendente, um grande número de métodos baseados em similaridade com diferentes definições de similaridade foram propostas. Foi demonstrado empiricamente que a similaridade entre nós pode ser definida em termos de propriedades topológicas da rede (MARTÍNEZ; BERZAL; CUBERO, 2016).

Para predição de *links*, cada elemento corresponde a um par de vértices com o rótulo indicando o status do *link*, portanto, as características escolhidas devem representar alguma forma de proximidade entre o par de vértices. Na pesquisa de predição de *links*, a maioria das características é extraída da topologia do grafo. Além disso, alguns trabalhos desenvolvem um conjunto de características construídas a partir de modelo de evolução de

grafos. Além desses, os atributos de vértices e arestas podem também ter características muito boas para muitos domínios de aplicações (HASAN; ZAKI, 2011).

As características baseadas na topologia de grafos são as mais naturais para predição de *links*. Aqui são chamadas de características topológicas. De fato, muitos trabalhos (LIBEN-NOWELL; KLEINBERG, 2007) (KASHIMA; ABE, 2006) na predição de *links* concentram-se apenas no grafo topológico do conjunto de características. Normalmente, eles calculam a similaridade com base nas vizinhanças dos nós ou nos conjuntos de caminhos entre um par de nós. A vantagem dessas características é que elas são genéricas e são aplicáveis a grafos de qualquer domínio. Portanto, nenhum conhecimento de domínio é necessário para calcular os valores desses recursos da rede complexa (HASAN; ZAKI, 2011).

2.3.2.1 Abordagens Locais

As abordagens baseadas em similaridade local usam informações estruturais relacionadas à vizinhança do nó para calcular a similaridade de cada nó com outros nós na rede. Estas abordagens são mais rápidas que as técnicas não-locais e altamente paralelizáveis. Além disso, elas nos permitem lidar eficientemente com o problema de predição de *links* de maneira muito dinâmica e em redes que estão em constante mudança, como redes sociais online (MARTÍNEZ; BERZAL; CUBERO, 2016).

Sua principal desvantagem é que o uso apenas de informações locais restringe o conjunto de nós de similaridade que pode ser calculado para distância de dois nós (vizinhos de vizinhos). Isso pode ser uma grande desvantagem, pois muitos *links* são formados a distâncias maiores que duas em muitas redes do mundo real, especialmente em redes do tipo que não são *small world* (LIBEN-NOWELL; KLEINBERG, 2007). No entanto, esses métodos demonstraram uma precisão de predição muito competitiva em relação a técnicas mais complexas (MARTÍNEZ; BERZAL; CUBERO, 2016).

Algumas das mais populares funções de similaridade amplamente utilizadas no estado da arte (WANG et al., 2015),(MARTÍNEZ; BERZAL; CUBERO, 2016), são as seguintes: *Vizinhos Comuns* (NEWMAN, 2001a), *Adamic-Adar* (ADAMIC; ADAR, 2003), *Ligação Preferencial* (NEWMAN, 2001a), *Similaridade de Jaccard* (LIBEN-NOWELL; KLEINBERG, 2007).

Vizinhos Comuns (VC): Vizinhos comuns é a técnica local mais simples. A semelhança entre dois nós é definida como o número de vizinhos compartilhados entre ambos os nós (LIBEN-NOWELL; KLEINBERG, 2007).

Faz sentido supor que, se duas pessoas compartilham muitos conhecidos, é mais provável que se encontrem do que dois indivíduos sem contatos comuns. Diferentes estudos confirmaram essa hipótese observando uma correlação entre o número de vizinhos compar-

tilhados entre pares de nós e a probabilidade de serem vinculados (NEWMAN, 2001a). Este método define a função de similaridade por meio da Equação 2.4.

$$s(x, y) =\mid \Gamma(x) \cap \Gamma(y) \mid \tag{2.4}$$

Apesar de sua simplicidade, essa medida tem um desempenho surpreendentemente bom. Este método é a base para outras abordagens apresentadas posteriormente.

Segundo (HASAN; ZAKI, 2011), para dois nós, x e y, o tamanho de seus vizinhos em comum é definido como $\mid \Gamma(x) \cap \Gamma(y) \mid$. A ideia de usar o tamanho de vizinhos comuns é apenas um atestado da propriedade de transitividade da rede. Em simples palavras, significa que nas redes sociais se o vértice x estiver conectado ao vértice z e vértice y está conectado ao vértice z, existe uma probabilidade aumentada de o vértice x também ser conectado ao vértice y.

Então, se o número de vizinhos comuns cresce, a chance de que x e y terão um *link* entre eles aumenta. Calculou-se essa quantidade no contexto de redes de colaboração para mostrar que existe uma correlação positiva entre o número de vizinhos comuns de x e y no tempo t, e a probabilidade de que eles irão colaborar no futuro (NEWMAN, 2001a).

Similaridade de Jaccard (JS): Esse coeficiente amplamente utilizado em sistemas de recuperação de informações foi proposto por (JACCARD, 1901) para comparar a similaridade e diversidade de conjuntos de amostras. Ele mede a proporção de vizinhos compartilhados no total do conjunto de vizinhos para dois nós. Essa função de similaridade é definida por meio da Equação 2.5.

$$s(x, y) = \frac{\mid \Gamma(x) \cap \Gamma(y) \mid}{\mid \Gamma(x) \cup \Gamma(y) \mid} \tag{2.5}$$

Pode-se ver facilmente que esse método é mais uma variação do método de vizinhos comuns. Há uma penalização para cada vizinho não compartilhado.

Como a métrica Vizinhos Comuns não é normalizada, então pode-se usar a Similaridade de Jaccard, que normaliza o tamanho de vizinhos. Conceitualmente, define a probabilidade de que um vizinho comum de um par de vértices x e y seriam selecionados se a seleção fosse feita aleatoriamente da união dos conjuntos vizinhos de x e y.

Então, para um alto número de vizinhos, a pontuação seria maior. No entanto, a partir dos resultados experimentais de quatro redes de colaboração diferentes, (LIBEN-NOWELL; KLEINBERG, 2007) mostraram que o desempenho da Similaridade de Jaccard é pior em comparação com o número de vizinhos comuns.

Adamic-Adar (AA): Essa medida de similaridade, proposta inicialmente por (ADAMIC; ADAR, 2003), pretendia medir a semelhança entre duas entidades com base em seus recursos compartilhados. No entanto, cada peso de recurso é penalizado logaritmica-

mente por sua frequência de aparência. Se considerarmos os vizinhos como características, pode ser escrito como a Equação 2.6.

$$s(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log \mid \Gamma(z) \mid} \tag{2.6}$$

Essa equação é uma variação da função de similaridade de vizinhos comuns em que cada vizinho compartilhado é penalizado por seu grau. Isso intuitivamente faz sentido em um grande número de redes reais. Por exemplo, nas redes sociais, a quantidade de recursos ou tempo que um nó pode gastar em cada um de seus vizinhos diminui à medida que seu grau aumenta, diminuindo também sua influência sobre eles.

Dessa forma, a métrica Adamic-Adar pesa os vizinhos comuns em menor grau mais fortemente. A partir dos resultados relatados dos trabalhos existentes sobre predição de links, o Adamic-Adar funciona melhor do que as métricas Vizinhos Comuns e Similaridade de Jaccard.

Ligação Preferencial (LP): Este índice é um resultado direto do conhecido modelo de formação de redes complexas Barabási-Albert (BARABÁSI; ALBERT, 1999) (MITZENMACHER, 2004). Albert-László Barabási e Réka Albert construíram um modelo teórico baseado na observação de que a probabilidade de formação de *links* entre dois nós aumenta com o grau desses nós.

A semelhança entre dois nós, de acordo com o modelo de Barabási-Albert, pode ser representada por meio da Equação 2.7

$$s(x,y) = \mid \Gamma(x) \mid\mid \Gamma(y) \mid \tag{2.7}$$

Essa medida também pode ser aplicada em contextos não-locais, uma vez que não depende de vizinhos compartilhados. No entanto, sua precisão de predição geralmente é baixa quando aplicada como uma medida global.

2.3.2.2 Abordagens Globais

Outro conjunto de abordagens para predição de *links* é baseado na agregação (pesada ou normalizada) de contagens de caminhos de comprimentos variados. Em contraste com os métodos locais descritos acima, esses métodos usam informações globais sobre toda a rede. Por exemplo, a similaridade Katz conta o número de caminhos entre dois nós, caminhos de ponderação de comprimento k por β^k . Outra classe de métodos globais são métodos baseados em difusões conservadoras tais como PageRank.

PageRank Semeado (RPR - Rooted PageRank): Segundo (MUTLU; OGHAZ, 2019), PageRank (PR) é a métrica usada pela Pesquisa do Google para determinar a

importância relativa das páginas da web, tratando os *links* como um voto. O valor RPR define que a classificação de um nó é proporcional à probabilidade de que ele possa ser alcançado de forma aleatória (BRIN; PAGE, 1998) (WANG et al., 2015). O valor do RPR é representado por meio da Equação 2.8. O valor do λ especifica a probabilidade de o algoritmo visitar os vizinhos do nó (WANG et al., 2015).

$$s_{(x,y)}^{RPR} = (1 - \lambda)(1 - \lambda P_{x,y})^{-1} \tag{2.8}$$

onde, $P_{i,j} = D^{-1}A$, onde A é a matriz de adjacência e D é uma matriz tal que se $i \neq j$, $D_{i,j} = 0$ e $D_{i,i} = \sum_{1 \leq j \leq n} A_{i,j}$. Nota-se que é possível calcular o valor de PR calculando a média das colunas do RPR (SONG et al., 2009).

Katz: Katz soma diretamente todos os caminhos que existem entre um par de vértices x e y (HASAN; ZAKI, 2011). Mas, para penalizar a contribuição de caminhos mais longos no cálculo de similaridade, amortece exponencialmente a contribuição de um caminho por um fator de β^l, onde l é o comprimento do caminho. O valor da similaridade de Katz para um par de vértices x, y é dado pela Equaçao 2.9.

$$katz(x, y) = \sum_{l=1}^{\infty} \beta^l . \mid paths_{x,y}^{(l)} \mid \tag{2.9}$$

onde $|paths_{x,y}^{(l)}|$ é o número de elementos do conjunto de todos os caminhos de comprimento l de x até y.

Katz baseia-se no conjunto de todos os caminhos entre os nós x e y. O parâmetro β (≤ 1) pode ser usado para regularizar esse recurso. Um pequeno valor de β considera apenas os caminhos mais curtos para o qual esse recurso se comporta muito como recursos baseados no vizinhança do nó. Um problema com esse recurso é que ele é computacionalmente caro. Pode-se mostrar que o *score* de Katz entre todos os pares de vértices pode ser calculado por $(I$ - $\beta A)^{-1} - I$, onde A é a matriz de adjacência e I é uma matriz de identidade de ordem n.

2.3.3 Estrutura de Ordem Mais Complexa

Como uma rede codifica relacionamentos entre pares (arestas) e entre elementos (nós), o problema de predição de *links* é natural em muitos casos. No entanto, estudos recentes demonstraram que as redes evoluem através de interações de ordem mais complexa, ou seja, a estrutura em redes em evolução envolve interações entre mais do que apenas dois nós. Pesquisas recentes também introduziram o problema de prever o momento em que uma adição de aresta irá fechar um triângulo.

Além disso, modelos de grafos aleatórios construídos a partir de distribuições de triângulos mostraram ser bons ajustes para dados, fornecendo evidências adicionais de que as relações triádicas são importantes para a montagem de redes.

2.3.4 Métodos Para Predição de Links Aos Pares

O artigo (NASSAR; BENSON; GLEICH, 2019a) propõe, inicialmente, usar novas versões de medidas de similaridade, que foram denominadas "predição de *links* aos pares". Primeiro, foram adaptados os métodos locais Vizinhos Comuns, Similaridade de Jaccard, Adamic-Adar e Ligação Preferencial para medir a similaridade entre os nós e as arestas. Nesse mesmo artigo, foram propostos métodos baseados em difusão a partir do PageRank Semeado, reforçando o PageRank pela utilização de triângulos. A seguir, descrevemos as medidas propostas em (NASSAR; BENSON; GLEICH, 2019a) e (NASSAR; BENSON; GLEICH, 2020).

Medidas Locais de Similaridade para a Predição Aos Pares

O objetivo foi estender métodos locais comuns para a predição de *links* para o cenário de predição de *link* aos pares. Em outras palavras, ao invés de calcular a similaridade entre nós, agora foi calculada a similaridade entre uma aresta e um nó. Para fazer isso, simplesmente substituíram os vizinhos de um nó pelos vizinhos de uma aresta. Isso requer que fosse especificado o que os vizinhos de uma aresta (u, v) deveriam capturar. As definições dos conjuntos $\Gamma^*((u, v))$ e $\Gamma^{**}((u, v))$ foram apresentadas na Seção 2.3.

Os autores utilizaram a versão do conjunto de vizinhos de uma aresta com a união dos vizinhos de cada vértice da aresta ($\Gamma^*((u, v))$) (NASSAR; BENSON; GLEICH, 2019b) (NASSAR; BENSON; GLEICH, 2020). Posteriormente, com a interseção dos vizinhos de cada vértice da aresta ($\Gamma^{**}((u, v))$) (NASSAR; BENSON; GLEICH, 2019a). Na definição usando interseção, o conjunto de vizinhos de uma aresta é mais restrito, comparado com o obtido usando união.

Então, foram definidas medidas de similaridade correspondentes à nova definição de vizinhança de uma aresta nas duas versões:

· Vizinhos Comuns aos pares (VC*), utilizando a métrica 'ou' é apresentada na Equação 2.10.

$$VC^*(w, (u, v)) = \mid \Gamma(w) \cap \Gamma^*((u, v)) \mid \tag{2.10}$$

· Vizinhos Comuns aos pares (VC**), utilizando a métrica 'e' é apresentada na Equação 2.11.

$$VC^{**}(w, (u, v)) = \mid \Gamma(w) \cap \Gamma^{**}((u, v)) \mid \tag{2.11}$$

· Similaridade de Jaccard aos pares (JS*), utilizando a métrica 'ou' é apresentada na Equação 2.12.

$$JS^{*}(w,(u,v)) = \frac{\mid \Gamma(w) \cap \Gamma^{*}((u,v)) \mid}{\mid \Gamma(w) \cup \Gamma^{*}((u,v)) \mid} \tag{2.12}$$

· Similaridade de Jaccard aos pares (JS**), utilizando a métrica 'e' é apresentada na Equação 2.13.

$$JS^{**}(w,(u,v)) = \frac{\mid \Gamma(w) \cap \Gamma^{**}((u,v)) \mid}{\mid \Gamma(w) \cup \Gamma^{**}((u,v)) \mid} \tag{2.13}$$

· Adamic–Adar aos pares (AA*), utilizando a métrica 'ou' é apresentada na Equação 2.14.

$$AA^{*}(w,(u,v)) = \sum_{z \in \Gamma(w) \cap \Gamma^{*}((u,v))} \frac{1}{log \mid \Gamma(z) \mid} \tag{2.14}$$

· Adamic–Adar aos pares (AA**), utilizando a métrica 'e' é apresentada na Equação 2.15.

$$AA^{**}(w,(u,v)) = \sum_{z \in \Gamma(w) \cap \Gamma^{**}((u,v))} \frac{1}{log \mid \Gamma(z) \mid} \tag{2.15}$$

· Ligação Preferencial aos pares (LP*, utilizando a métrica 'ou') é apresentada na Equação 2.16.

$$LP^{*}(w,(u,v)) = \mid \Gamma(w) \mid \cdot \mid \Gamma^{*}((u,v)) \mid \tag{2.16}$$

· Ligação Preferencial aos pares (LP**), utilizando a métrica 'e' é apresentada na Equação 2.17.

$$LP^{**}(w,(u,v)) = \mid \Gamma(w) \mid \cdot \mid \Gamma^{**}((u,v)) \mid \tag{2.17}$$

Os autores (NASSAR; BENSON; GLEICH, 2019a) desenvolveram uma nova métrica para a predição de *link* aos pares baseado no PageRank Semeado que será apresentada a seguir.

PageRank Semeado aos Pares

O PageRank Semeado foi proposto como um método que permite analisar o fluxo de informações em uma rede, com o objetivo de prever *links* e comunidades (ANDERSEN; CHUNG; LANG, 2006) (GLEICH, 2015). PageRank Semeado modela o fluxo de informação do nó semeado para outros nós na rede através de uma cadeia de Markov e a distribuição

estacionária da cadeia fornece as pontuações nos nós. Uma pontuação alta em um nó é um sinal de que o nó deve estar conectado ao nó semente.

Mais formalmente, A é a matriz de adjacência de um grafo não direcionado e P é a matriz estocástica por coluna de um passeio aleatório nesse grafo. Especificamente, $P_{i,j} = A_{i,j}/ \mid \Gamma(j) \mid$. Dado o nó u como o nó semente. As pontuações do PageRank são calculadas como entradas do vetor x, solução do sistema linear $(I$ - αP $)x = (1$ - $\alpha)$ e_u, onde e_u é o vetor de todos os zeros, exceto no índice correspondente ao nó u, o qual é igual a 1 (ou seja, e_u é o vetor indicador no nó u). O parâmetro α é a probabilidade de transição de acordo com a distribuição de probabilidade em P e (1 - α) é a probabilidade de teletransporte de acordo com a distribuição de probabilidade em e_u.

As entradas do vetor x fornecem similaridades entre o nó u e os outros nós do grafo e, portanto, podem ser usadas para a predição de *links* tradicional. Da mesma forma que o PageRank Semeado prevê a relevância de outros nós na rede para um único nó origem, foi proposto o **PageRank Semeado aos Pares** para prever a relevância dos nós para uma única aresta dada. Com essas similaridades, é possível fazer predições para o problema de predição de *links* aos pares. Para uma determinada aresta (u, v), o PageRank Semeado aos Pares resolve o seguinte sistema linear:

$(I$ - $\alpha P)x = (1$ - $\alpha)e_{u,v}$.

Onde $e_{u,v}$ é o vetor de todos os zeros, exceto nos índices u e v, que valem 1/2. A solução x pode ser interpretada como a similaridade de cada nó para a aresta (u, v).

O PageRank Semeado aos Pares é equivalente à soma do PageRank Semeado único em cada um dos nós até um múltiplo escalar. Isso segue rapidamente da linearidade do problema do PageRank. Para ver isso, $\overline{x}_u$ e $\overline{x}_v$ são os valores solução de PageRank Semeado correspondentes aos nós u e v respectivamente. Então,

$(I$ - αP $)\overline{x}_u = (1$ - $\alpha)e_u$

$(I$ - αP $)\overline{x}_v = (1$ - $\alpha)e_v$

Adicionando as duas equações anteriores:

$(I$ - αP $)(\overline{x}_u + \overline{x}_v) = (1$ - $\alpha)(e_u + e_v)$

$(I$ - αP $)(\overline{x}_u + \overline{x}_v) = (1$ - $\alpha)(2e_{u,v})$

$1/2(I$ - αP $)(\overline{x}_u + \overline{x}_v) = (1$ - $\alpha)(e_{u,v})$

$(I$ - $\alpha P)x = (1$ - $\alpha)e_{u,v}$

Assim, $2x = \overline{x}_u + \overline{x}_v$, e a solução de PageRank Semeado aos Pares é equivalente ao somatório das equações do PageRank com uma única semente até o escalonamento. Observar que a ideia de semear aos pares o PageRank pode ser estendida para mais de dois nós (LOFGREN; BANERJEE; GOEL, 2016).

3 TRABALHOS RELACIONADOS

3.1 Predição de *links*

Segundo (OTTE; ROUSSEAU, 2002), a análise de redes sociais não é uma teoria formal em sociologia, mas uma estratégia para investigar estruturas sociais. Como é uma ideia que pode ser aplicada em muitos campos, estudou-se, em particular, sua influência nas ciências da informação.

Os cientistas da informação estudam redes de publicação, citação e cocitação, estruturas de colaboração e outras formas das redes de interação social. Além disso, a internet representa uma rede social de uma escala sem precedentes. A análise de redes sociais está mais relacionada às teorias sobre a economia de livre mercado, geografia e redes de transporte.

Em 2007, (LIBEN-NOWELL; KLEINBERG, 2007) fizeram a seguinte pergunta: dado um nó em uma rede social, pode-se inferir quais novas interações entre seus membros provavelmente ocorrerão no futuro próximo?

Assim, formalizou-se essa questão como o problema de predição de *link* e desenvolveram-se abordagens para vincular predição baseada em medidas para analisar a "proximidade" de nós em uma rede. Experimentos em grandes redes de coautoria sugerem que informações sobre futuras interações podem ser extraídas apenas da topologia de rede, e que medidas bastante sutis para detectar a proximidade do nó podem superar as medidas mais diretas.

Algumas predições representativas de *links* foram pesquisadas por (HASAN; ZAKI, 2011). Essas predições foram categorizadas pelo tipo de modelo. Três tipos de modelos foram amplamente considerados: primeiro, os modelos tradicionais (não bayesianos) que extraem um conjunto de recursos para treinar um modelo de classificação binária. Segundo, o probabilístico com abordagens que modelam a probabilidade conjunta entre as entidades em uma rede por modelos gráficos bayesianos. E, finalmente, a abordagem baseada em álgebra linear que calcula a semelhança entre os nós em uma rede por matrizes de similaridade com classificação reduzida.

Em 2012, introduziu-se o conceito de um perfil de colocação de vértices (VCP) para fins de análise e predição de *links* topológicos (LICHTENWALTER; CHAWLA, 2012). Os VCPs fornecem informações quase completas sobre a estrutura local circundante de pares de vértices incorporados.

A abordagem VCP oferece uma nova ferramenta para especialistas em domínio compreenderem os mecanismos subjacentes de crescimento de redes e analisarem os mecanismos de formação de ligações nos contextos sociológico, biológico, físico ou outro

apropriado.

A mesma resolução que dá à VCP seu poder de capacidade analítica, também permite um bom desempenho quando usado em modelos supervisionados para discriminar possíveis novos *links*. Os métodos VCP foram demonstrados executando a predição competitivamente com métodos não supervisionados e supervisionados em várias famílias de redes diferentes.

Para resolver o problema de predição de *links*, (RÜMMELE; ICHISE; WERTHNER, 2015) seguiram a abordagem de contar *graphlets* de 3 nós, que são subgrafos induzidos de um grafo G de 3 vértices, e sugeriram três extensões para o método original. Ao realizar experimentos em duas redes sociais reais, mostraram que os novos métodos têm um poder preditivo, no entanto, a evolução da rede não pode ser explicada por um recurso específico em todos os momentos.

Observaram também que algumas propriedades de rede podem apontar para recursos mais eficazes para a predição de *link* temporal.

Em 2019, (MUTLU; OGHAZ, 2019) analisaram o objetivo geral das técnicas do problema de predição de *links*. Foi o primeiro estudo que considerou todos os desafios sobre o estudo de redes e sua abordagem através dos modelos de aprendizado de máquina.

Contudo, em 2019 (NASSAR; BENSON; GLEICH, 2019a) identificaram que a evolução da rede é frequentemente mediada por estruturas de ordens mais complexas envolvendo mais do que pares de nós. Por exemplo, subgrafos completos de três nós (também chamados triângulos) são fundamentais para a estrutura das redes sociais, mas a estrutura tradicional de predição de *links* não prevê diretamente essas estruturas.

Para atender a essa necessidade, os autores propuseram uma nova tarefa de predição de *link* chamada predição de *link* 'aos pares' que tem como objetivo fazer a predição de novos triângulos, com a finalidade de encontrar os nós que provavelmente formarão um triângulo com uma aresta.

3.2 Redes heterogêneas e métricas topológicas

Os autores de (HUANG; LI; CHEN, 2005) propõem uma adaptação das métricas 'tradicionais' em redes homogêneas para serem utilizadas em redes heterogêneas bipartidas, isto é, uma rede na qual os nós são de dois tipos diferentes (uma bipartição do conjunto de nós) e todas as arestas têm extremidades em conjuntos diferentes da bipartição. Os autores propõem transformar o conjunto $\Gamma(u)$ em $\hat{\Gamma}(u) = \cap_{v \in \Gamma(u)} \Gamma(v)$ (vizinhos dos vizinhos do vértice u).

No artigo de Liben-Nowell e Kleinberg (LIBEN-NOWELL; KLEINBERG, 2007), trabalho importante em predição de *links*, é analisada uma rede de coautoria acadêmica

utilizando características topológicas da rede para prever a formação de arestas entre dois nós não conectados. Neste caso, a rede de coautoria é homogênea, ou seja, todos os nós são do mesmo tipo.

Em 2010, (BENCHETTARA; KANAWATI; ROUVEIROL, 2010) apresentou uma abordagem diferente da proposta em (HUANG; LI; CHEN, 2005) para tratar redes heterogêneas bipartidas. Neste caso, é utilizada uma projeção do grafo que representa a rede sobre um dos dois conjuntos da bipartição e definidas as métricas de acordo com essa projeção.

A Figura 8 representa as projeções de um grafo bipartido, segundo (BENCHETTARA; KANAWATI; ROUVEIROL, 2010).

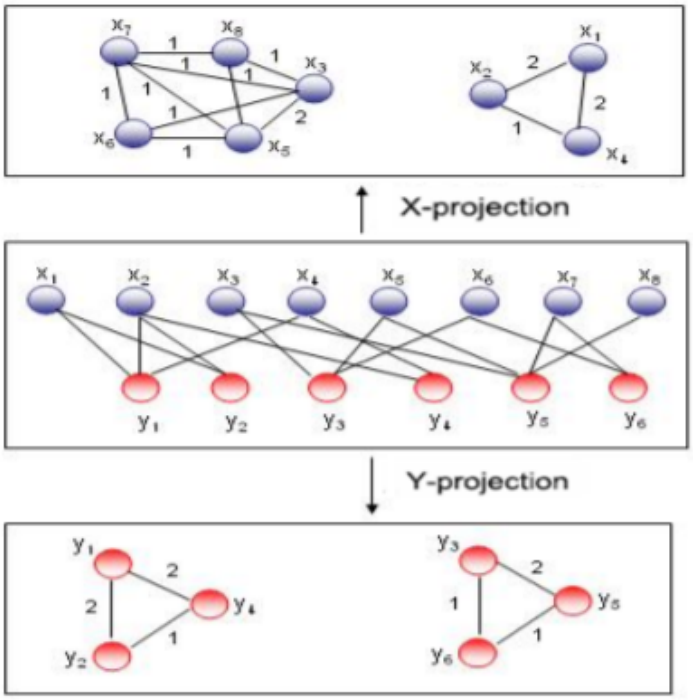

Figura 8 – Projeções de um grafo bipartido (BENCHETTARA; KANAWATI; ROUVEIROL, 2010).

Em (DAVIS; LICHTENWALTER; CHAWLA, 2013) foi desenvolvida uma abordagem para uma rede heterogênea multipartida. O método proposto, denominado MRLP (*multi-relational link prediction*), cujo componente principal é utilizar um esquema de pesos para diferentes tipos de combinações de arestas, a partir da contagem de subgrafos formados por 3 nós que existem na rede.

Já em 2019, Nassar et al. (NASSAR; BENSON; GLEICH, 2019a) propõem prever a formação de arestas considerando um nó e uma aresta existente na rede.

Os artigos (LIBEN-NOWELL; KLEINBERG, 2007) e (NASSAR; BENSON; GLEICH, 2019a) consideram de formas diferentes o fechamento de triângulos ao analisar subgrafos

formados por 3 nós em redes homogêneas (ver Figura 9 e Figura 10).

Figura 9 – O primeiro grafo representa G_t, as ligações existentes até o instante t, e o segundo $G_{t'}$, as ligações existentes até o instante t', para $t < t'$. A aresta tracejada seria uma possível solução usando métricas para predição de *links* tradicional, para o par de vértices não adjacentes x e y. Adaptado da Figura 1 (NASSAR; BENSON; GLEICH, 2019a).

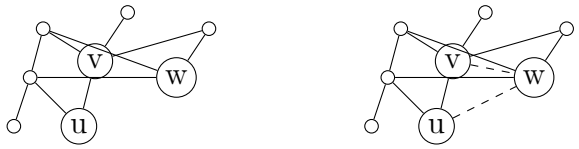

Figura 10 – O primeiro grafo representa G_t, as ligações existentes até o instante t, e o segundo $G_{t'}$, as ligações existentes até o instante t', para $t < t'$. As arestas tracejadas seriam possíveis soluções usando métricas para predição de *links* aos pares para o vértice w e a aresta (u, v). Adaptado da Figura 1 (NASSAR; BENSON; GLEICH, 2019a).

3.3 Conclusões

A Tabela 1 representa os trabalhos relacionados contendo os tipos de informações da rede, o tipo da predição de *links* ('tradicional' e 'aos pares') e a solução referente às redes heterogêneas e homogêneas.

	Tipos de Informações			Predição de Links		Tipos de Redes	
Referência	**Topológicas**	**Contextuais**	**Temporais**	**Tradicional**	**Aos Pares**	**Homogêneas**	**Heterogêneas**
(OTTE; ROUSSEAU, 2002)	x			x			
(HUANG; LI; CHEN, 2005)	**x**			**x**		**x**	**x**
(LIBEN-NOWELL; KLEINBERG, 2007)	**x**			**x**		**x**	
(BENCHETTARA; KANAWATI; ROUVEIROL, 2010)	**x**			**x**		**x**	**x**
(HASAN; ZAKI, 2011)	x			x			
(LICHTENWALTER; CHAWLA, 2012)	x		x	x			
(DAVIS; LICHTENWALTER; CHAWLA, 2013)	**x**			**x**			**x**
(RÜMMELE; ICHISE; WERTHNER, 2015)	x	x	x	x			
(MUTLU; OGHAZ, 2019)	x			x			
(NASSAR; BENSON; GLEICH, 2019a)	**x**				**x**	**x**	
(NASSAR; BENSON; GLEICH, 2020)	**x**				**x**	**x**	

Tabela 1 – Tabela comparativa de trabalhos relacionados.

Até o nosso conhecimento, ainda não existe uma análise comparativa das métricas topológicas locais nas duas versões: tradicional e aos pares. Assim, este trabalho propõe fazer uma análise comparativa dos resultados obtidos ao utilizar as métricas topológicas locais tradicionais e aos pares em redes de coautoria. As redes homogêneas foram escolhidas para fazer os experimentos, pois são mais adequadas para fazer a comparação proposta.

4 METODOLOGIA

Neste capítulo, será apresentada a metodologia utilizada para fazer a análise comparativa das métricas topológicas nas duas versões ('tradicional' e 'aos pares').

4.1 Abordagem Proposta

Neste trabalho, considerou-se um grafo no qual cada aresta e de $G = (V, E)$ representa uma interação entre dois nós u e v num instante de tempo $t(e)$. Não serão consideradas múltiplas interações entre u e v. Para um dado instante de tempo t, nota-se por G_t o subgrafo de G que contém todas as arestas e tal que $t(e) < t$. A formulação matemática do problema é dada a seguir.

Dois instantes de tempo $t < t'$ foram escolhidos e foi considerado um algoritmo que acesse o grafo que representa a rede até o instante t, G_t, e que retorne uma lista de pares de elementos (dois nós não adjacentes, ou um nó e uma aresta em G_t) que são predições de arestas para $G_{t'}$. Os intervalos $(0, t]$ e $(t, t']$ são referidos como intervalo de treino e teste, respectivamente. Cada preditor p considerado retorna uma lista ordenada L_p de pares em $V \times V$, que são as predições de novas interações em $G_{t'}$ em ordem não crescente de confiança.

As redes usadas nos experimentos são redes de coautoria. O conjunto *Core* foi definido para consistir em todos os autores que escreveram pelo menos 1 artigo durante o período de treinamento e pelo menos 1 artigo durante o período de teste. Ou seja, considerou-se todos os autores que tiveram ao menos uma publicação entre eles ($Core = V$).

A medida de desempenho para o preditor p foi determinada da seguinte forma: escolheram-se os primeiros k pares de predições de novas interações da lista ordenada L_p (Top-k). Assim, para poder comparar cada métrica nas versões 'tradicional' e 'aos pares', selecionaram-se os primeiros k elementos da lista L_p. Finalmente, para cada L_p, determinaram-se as medidas de qualidade de classificação (precisão, acurácia, revocação e F-1). As medidas de qualidade citadas serão detalhadas na Subseção 4.1.1. Para se chegar aos resultados destas medidas, foi necessário gerar a matriz de confusão, que é apresentada na Tabela 2.

Classes	**Predita** C_+	**Predita** C_-
Verdadeira C_+	Verdadeiros Positivos	Falsos Negativos
Verdadeira C_-	Falsos Positivos	Verdadeiros Negativos

Tabela 2 – Matriz de Confusão de um Classificador - problema com 2 classes (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Além das medidas de qualidade mencionadas, também foram calculados o preditor randômico, e para a rede, a transitividade e o coeficiente de agrupamento médio. O preditor randômico será detalhado na Subseção 4.1.2. A transitividade e o coeficiente de agrupamento médio estão descritos no Capítulo 2, na Seção 2.2, Equações 2.1 e 2.3, respectivamente.

4.1.1 Medidas de qualidade

A avaliação dos resultados obtidos na tarefa de predição de *links* é realizada a partir de várias medidas de qualidade. As medidas que se destacam na literatura e utilizadas neste trabalho serão apresentadas a seguir.

Precisão

A precisão é a fração onde o numerador é o número de positivos classificados corretamente, conhecidos como verdadeiros positivos (VP) e o denominador é o número total que são classificados como positivo, conhecidos como verdadeiros positivos (VP) e falsos positivos (FP) (ZHANG; YU, 2014). A precisão está representada pela Equação 4.1.

$$\text{precisão} = \frac{VP}{VP + FP} \tag{4.1}$$

Acurácia

A acurácia é a razão onde o numerador é o número de instâncias classificadas corretamente no conjunto de teste, conhecidos como verdadeiros positivos (VP) e verdadeiros negativos (VN) e onde o denominador é o número total de instâncias, ou seja, todos os pares de vértices possíveis de se conectarem, representados por verdadeiros positivos (VP), falsos positivos (FP), falsos negativos (FN) e verdadeiros negativos (VN) (ZHANG; YU, 2014). A acurácia está representada por meio da Equação 4.2.

$$\text{acurácia} = \frac{VP + VN}{VP + FP + FN + VN} \tag{4.2}$$

Recall (Revocação ou Cobertura)

A medida recall é a fração onde o numerador é o número de arestas corretamente classificadas (VP) e o denominador é o número total de arestas reais no conjunto de teste, ou seja, a soma dos verdadeiros positivos e os falsos negativos (ZHANG; YU, 2014). O recall está representado por meio da Equação 4.3.

$$\text{recall} = \frac{VP}{VP + FN} \tag{4.3}$$

F-Mesuare (F-1)

A medida F-Mesuare (F-1) é a média harmônica entre as medidas precisão e recall (ZHANG; YU, 2014). O F-1 está representado por meio da Equação 4.4.

$$\text{F-1} = \frac{2\ .(\ \text{precisão}\ .\ \text{recall})}{\text{precisão} + \text{recall}} \tag{4.4}$$

4.1.2 Preditor Randômico

O preditor randômico é usado como base de comparação dos valores obtidos pelos preditores nos experimentos. E faz a predição pela seleção aleatória de um par de autores que não tenham colaborado em G_t. Notando $E_{new} = E_{t'} - E_t$ e $E_{old} = E_t$, o preditor randômico é apresentado na Equação 4.5. Como foi mencionado anteriormente, nos experimentos utilizamos $Core = V$.

$$\text{preditor randômico} = \frac{|E_{new}|}{|(Core\ \text{x}\ Core)| - |E_{old}|} \tag{4.5}$$

4.2 Descrição das métricas

As métricas utilizadas nos experimentos foram definidas na Seção 2.3.2.1 (versão tradicional) e na Seção 2.3.4 (versões aos pares 'ou' e 'e'). As Tabelas 3, 4 e 5 resumem essas fórmulas.

Tabela 3 – Métricas utilizadas nos experimentos para a versão tradicional.

Tradicional	
$VC(x,y) =$	$\lvert\Gamma(x) \cap \Gamma(y)\rvert$
$JS(x,y) =$	$\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$
$AA(x,y) =$	$\sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{log\lvert\Gamma(z)\rvert}$
$LP(x,y) =$	$\lvert\Gamma(x)\rvert\lvert\Gamma(y)\rvert$

Tabela 4 – Métricas utilizadas nos experimentos para a versão aos pares 'ou'.

Aos Pares (ou)	
$VC^*(w,(u,v)) =$	$\lvert\Gamma(w) \cap \Gamma^*((u,v))\rvert$
$JS^*(w,(u,v)) =$	$\frac{\lvert\Gamma(w)\cap\Gamma^*((u,v))\rvert}{\lvert\Gamma(w)\cup\Gamma^*((u,v))\rvert}$
$AA^*(w,(u,v)) =$	$\sum_{z\in\Gamma(w)\cap\Gamma^*((u,v))} \frac{1}{log\lvert\Gamma(z)\rvert}$
$LP^*(w,(u,v)) =$	$\lvert\Gamma(w)\rvert\lvert\Gamma^*((u,v))\rvert$

Tabela 5 – Métricas utilizadas nos experimentos para a versão aos pares 'e'.

Aos Pares (e)	
$VC^{**}(w,(u,v)) =$	$\lvert\Gamma(w) \cap \Gamma^{**}((u,v))\rvert$
$JS^{**}(w,(u,v)) =$	$\frac{\lvert\Gamma(w)\cap\Gamma^{**}((u,v))\rvert}{\lvert\Gamma(w)\cup\Gamma^{**}((u,v))\rvert}$
$AA^{**}(w,(u,v)) =$	$\sum_{z\in\Gamma(w)\cap\Gamma^{**}((u,v))} \frac{1}{log\lvert\Gamma(z)\rvert}$
$LP^{**}(w,(u,v)) =$	$\lvert\Gamma(w)\rvert\lvert\Gamma^{**}((u,v))\rvert$

4.3 *Datasets*

Para determinar o *dataset* a ser utilizado nos experimentos, observou-se que tanto a métrica tradicional quanto aos pares não produzem bons resultados em redes hetorogêneas bipartidas. Experimentos preliminares mostraram que a adaptação proposta em (HUANG; LI; CHEN, 2005) não funciona com as métricas aos pares. Portanto, escolheu-se a rede de coautoria, uma rede homogênea. A rede de coautoria pode ser pensada como a projeção no conjunto dos autores da rede heterogênea bipartida que relaciona autores e artigos proposta em (BENCHETTARA; KANAWATI; ROUVEIROL, 2010). Dessa forma, as redes deste trabalho são homogêneas, utilizando grafos não direcionados. Os vértices representam os autores e as arestas representam as publicações conjuntas.

Para fins de avaliar a estrutura dos *datasets* utilizados nos experimentos, usaremos os parâmetros transitividade e coeficiente de agrupamento médio definidos no Capítulo 2 (nas Equações 2.1 e 2.3).

De acordo com a descrição do tipo de rede na qual é possível utilizar as métricas tradicional e aos pares, foram escolhidos dois tipos de *datasets* (1 do Lattes e 5 do *ArXiv*) correspondentes a redes de coautoria para realizar os experimentos comparativos. O primeiro experimento foi com um *dataset* menor para analisar melhor os resultados para posteriormente fazer um segundo experimento com *datasets* maiores. Os *datasets* serão descritos a seguir.

Em (BARBOSA et al., 2011) foi criada uma rede de coautoria a partir de dados retirados da Plataforma Lattes do Conselho Nacional de Pesquisa (CNPq) em 11/10/2010. Os nós do grafo representam autores e as arestas representam pelo menos 1 publicação conjunta entre dois autores (grafo não direcionado, sem arestas múltiplas).

Posteriormente, a rede criada anteriormente foi atualizada com dados até 10/10/2014 (MAGNANI, 2015). As Figuras 11 e 12 abaixo ilustram os grafos G_{2011} e G_{2014} que representam as duas redes de coautoria mencionadas anteriormente, considerando publicações conjuntas dos autores até o ano 2011 e até o ano 2014, respectivamente.

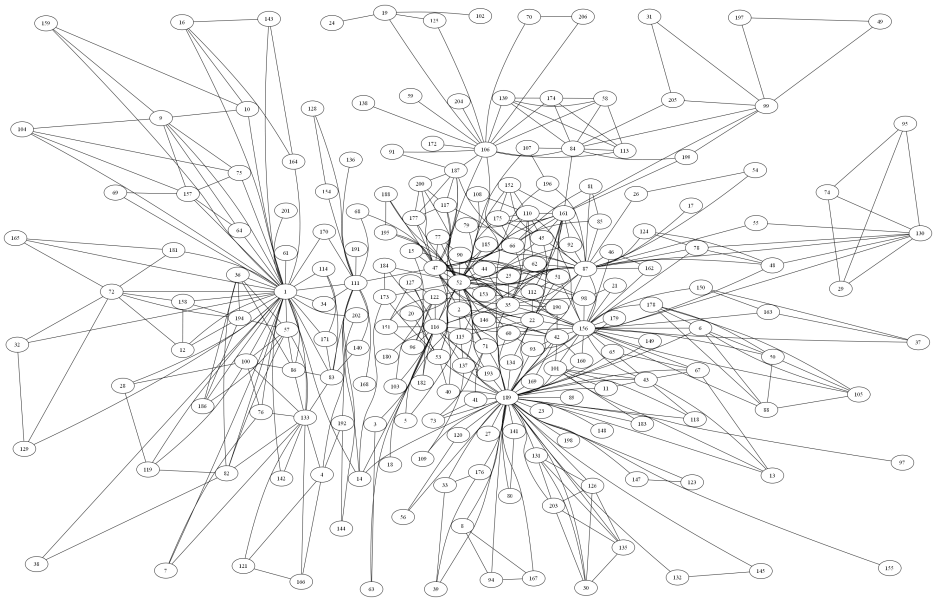

Figura 11 – G_{2011} (BARBOSA et al., 2011).

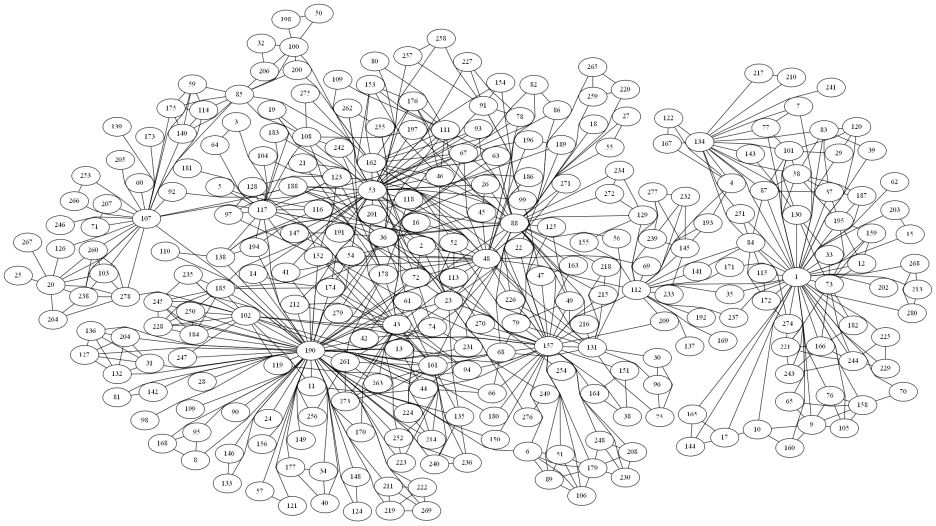

Figura 12 – G_{2014} (MAGNANI, 2015).

Além da rede Lattes, os grafos de coautoria utilizados neste trabalho também foram extraídos das seguintes redes de coautoria disponíveis no repositório *ArXiv*[1]:

- Astro-ph (Astrofísica)
- Cond-mat (Matéria Condensada)
- Gr-qc (Relatividade geral e cosmologia quântica)
- Hep-ph (Física em energia alta - Fenomenologia)
- Hep-th (Física em energia alta - Teoria)

ArXiv é um repositório de *pre-prints* em versão digital nos campos de matemática, física, astronomia, computação científica, biologia quantitativa, estatística, finanças quantitativas, etc. *Datasets* com informações de *pre-prints* em 5 campos de astrofísica foram usados em diversos trabalhos para avaliar os resultados de métodos de predição de *links*, como por exemplo (NEWMAN, 2001b), (LESKOVEC; KLEINBERG; FALOUTSOS, 2007), (PUJARI, 2015) , (CAVALCANTE; JUSTEL; GOLDSCHMIDT, 2020).

Neste trabalho, utilizaremos *pre-prints* do período 1992-1998 correspondentes aos *datasets* *General Relativity and Quantum Cosmology* (gr-qc), *Condense Matter Physics* (cond-mat), *Astro Physics* (astro-ph), *High Energy Physics Phenomenology* (hep-ph) e *High Energy Physics Theory* (hep-th). Os nós do grafo destes 5 *datasets* representam autores e as arestas representam pelo menos 1 publicação conjunta entre dois autores (grafo não direcionado, sem arestas múltiplas).

Para os cinco *datasets* do *ArXiv*, utilizaram-se algumas heurísticas para tratar alguns problemas eventuais. Percebeu-se que alguns autores eram os mesmos, mas, por algum motivo, foram citados com os sobrenomes de maneiras diferentes nos artigos. Para tentar minimizar esse problema, os nomes dos autores foram construídos a partir da inicial do primeiro nome e a inicial do último sobrenome.

4.3.1 Definição dos *Datasets*

A Tabela 6 apresenta os tamanhos dos conjuntos de nós e arestas dos grafos G_{2011} e G_{2014}, com informações correspondentes aos períodos $(2010, 2014]$.

A Tabela 7 apresenta os tamanhos dos conjuntos de nós e arestas dos grafos G_{1997} com informações correspondentes aos períodos $[1992, 1997]$ e as novas arestas do ano de 1998, respectivamente, correspondentes aos cinco *datasets* do *ArXiv*.

[1] <https://arxiv.org/>

Tabela 6 – Informações das redes de coautoria obtidas da Plataforma Lattes - CNPq. V_{2011} e E_{2011} são os conjuntos de autores e publicações até 2011. E_{2014} é o conjunto de publicações até 2014 dos autores em V_{2011}.

Dataset	$\|V_{2011}\|$	$\|E_{2011}\|$	$\|E_{2014} - E_{2011}\|$	*C(G)*	*T(G)*
$\boldsymbol{G}_{2011}$	207	520	236	0,7458	0,2137

Tabela 7 – Informações das redes de coautoria do *ArXiv*. V_{1997} e E_{1997} são os conjuntos de autores e publicações até 1997. E_{1998} é o conjunto de publicações até 2018 dos autores em V_{1997}.

Dataset	$\|V_{1997}\|$	$\|E_{1997}\|$	$\|E_{1998} - E_{1997}\|$	*C(G)*	*T(G)*
gr-qc	2.621	5.528	394	0,5009	0,7402
cond-mat	8.354	20.526	2.398	0,6064	0,4076
astro-ph	8.073	47.604	10.123	0,6658	0,5510
hep-ph	6.760	32.973	2.306	0,5807	0,6280
hep-th	6.238	12.832	1.215	0,4692	0,3055

4.4 Ambiente de apoio à experimentação

Com o objetivo de validar o método proposto por este trabalho, foi implementado um código na liguagem de programação Python (LUTZ, 1996) (versão 3.9.4), uma linguagem muito usada nas aplicações científicas e foi utilizada a biblioteca *NetworkX* (HAGBERG; SCHULT; SWART, 2008) (versão 2.5.1). Esta biblioteca é aplicada para a criação e manipulação de grafos.

Os experimentos da Plataforma Lattes foram realizados em um computador Intel Core i7, CPU 1.80GHz, 8GB de RAM. A execução dos experimentos dos *datasets* do *ArXiv* ocorreu em um ambiente computacional contendo o sitema operacional Ubuntu 20.04.2 LTS, com 8 núcleos de processador e 128 Gigabytes de memória RAM, uma máquina integrante do Laboratório de Computação de Alto Desempenho - Defesa Cibernética do IME.

5 EXPERIMENTOS E RESULTADOS

5.1 Considerações iniciais

Neste capítulo são apresentados dois experimentos utilizados para avaliar o método proposto, além de analisar os resultados obtidos nestes experimentos.

Os experimentos foram feitos em duas partes, a primeira para comparar a métrica tradicional com a versão aos pares 'ou', denominados Experimento 1.1 e Experimento 2.1 e a segunda para comparar a métrica tradicional com a métrica aos pares 'e', denominados Experimento 1.2 e Experimento 2.2. O Experimento 1 refere-se à rede de coautoria Lattes e o Experimento 2 compreende as cinco redes de coautoria do *ArXiv*.

5.2 Experimento 1

5.2.1 Descrição do experimento 1

O primeiro experimento foi realizado na rede de coautoria Lattes, considerando os grafos G_t e $G_{t'}$, para t, t', $t = 2011$ e $t' = 2014$. Observou-se que em G_{2014} existem 32 arestas que não pertencem ao grafo G_{2011}. O experimento está dividido em duas partes, a primeira para comparar a métrica aos pares na versão 'ou' (Experimento 1.1) e a segunda para a comparação com a métrica aos pares na versão 'e' (Experimento 1.2).

No Experimento 1.1, foram considerados valores de Top-k, para k = 3,5,7. Como ocorreram pequenas diferenças dos valores de Top-k obtidos com os k considerados, são mostrados unicamente os resultados para $k = 7$. O valor $k = 7$ foi escolhido porque a métrica tradicional Vizinhos Comuns apresentou unicamente 9 valores de *scores* diferentes.

No Experimento 1.2, dependendo da métrica, foram considerados diferentes valores de Topk-k para poder obter valores de Verdadeiro Positivo maiores que zero. Para a métrica Vizinhos Comuns, o preditor tradicional retornou 6 valores diferentes de *score* e, portanto, $k = 3$ foi escolhido.

Para Similaridade de Jaccard, foi considerado $k = 5$, pois os *scores* retornaram em maior quantidade comparados a Vizinhos Comuns. Para esta métrica, não foi considerado o $k = 7$, pois, ainda assim, as medidas de Precisão, F-1 e Revocação continuavam dando zero como resultado para a métrica Similaridade de Jaccard tradicional.

Já para as métricas Ligação Preferencial e Adamic-Adar foi considerado o $k = 7$. Para a métrica Adamic-Adar, com esse Top foi possível obter pelo menos uma aresta, que apareceu em G_{2014}, como Verdadeiro Positivo, tornando, assim, Precisão, F-1 e Revocação

diferentes de zero.

Em todos os casos anteriormente descritos, os valores de k para calcular o Top-k foram escolhidos após realizar testes para obter valores de VP (verdadeiros positivos) diferentes de zero.

5.2.2 Resultados obtidos do experimento 1

Todos os valores das medidas de qualidade de classificação dos *links* preditos obtidos do Experimento 1.1 são apresentados na Tabela 8. Pode-se concluir que, para Vizinhos Comuns e Similaridade de Jaccard, os resultados obtidos pelo método aos pares foram melhores ou iguais, exceto na Revocação para Vizinhos Comuns.

Para Ligação Preferencial, o método tradicional teve melhor resultado em todas as medidas de qualidade de classificação. E para Adamic-Adar, os resultados são balanceados para os dois métodos. Além disso, observamos que o preditor randômico obteve valor menor para as métricas tanto tradicional quanto aos pares no cálculo da Precisão, quando a mesma foi diferente de zero.

Tabela 8 – Resultados para G_{2011}, G_{2014} na rede Lattes do Experimento 1.1

Top-7			
Pred Rand	Medidas de	Métrica	
0,0015	qualidade	VC	VC^*
	Precisão	0,01010101	**0,025**
	Acurácia	0,965818951	**0,994732053**
	F-1	0,019310345	**0,035087719**
	Revocação	**0,21875**	0,058823529
		JS	JS^*
	Precisão	0	0
	Acurácia	0,985241094	**0,992497372**
	F-1	0	0
	Revocação	0	0
		LP	LP^*
	Precisão	**0,142857143**	0
	Acurácia	**0,998221239**	0,998029793
	F-1	**0,051282051**	0
	Revocação	**0,03125**	0
		AA	AA^*
	Precisão	**0,142857143**	0,058823529
	Acurácia	**0,998221239**	0,996928246
	F-1	0,051282051	**0,058823529**
	Revocação	0,03125	**0,058823529**

Todos os valores das medidas de qualidade de classificação dos *links* preditos obtidos

do Experimento 1.2 são apresentados na Tabela 9. Pode-se concluir que, para Vizinhos Comuns, os resultados obtidos pelo método aos pares foi melhor unicamente na Revocação.

Para Similaridade de Jaccard, o resultado foi melhor aos pares em todas as medidas de qualidade, exceto na Acurácia. Para Ligação Preferencial e Adamic-Adar, o método tradicional teve melhor resultado em todos os casos. Além disso, observamos que o preditor randômico obteve valor menor para as métricas tanto tradicional quanto aos pares no cálculo da Precisão, quando a mesma foi diferente de zero.

Tabela 9 – Resultados para G_{2011}, G_{2014} na rede Lattes do Experimento 1.2

Top-3			
Pred Rand	Medidas de	Métrica	
0,0015	qualidade	VC	VC^{**}
	Precisão	**0,333333333**	0,008368201
	Acurácia	**0,998413538**	0,987
	F-1	**0,057142857**	0,014652015
	Revocação	0,03125	**0,058823529**
Top-5			
		JS	JS^{**}
	Precisão	0	**0,002610966**
	Acurácia	**0,985625691**	0,94
	F-1	0	**0,005067568**
	Revocação	0	**0,085714286**
Top-7			
		LP	LP^{**}
	Precisão	**0,142857143**	0,1
	Acurácia	**0,998221239**	0,998029
	F-1	**0,051282051**	0,046511628
	Revocação	**0,03125**	0,03030303
Top-7			
		AA	AA^{**}
	Precisão	**0,142857143**	0,076923077
	Acurácia	**0,998221239**	0,99788
	F-1	**0,051282051**	0,043478261
	Revocação	**0,03125**	0,03030303

5.3 Experimento 2

5.3.1 Descrição do experimento 2

Neste experimento, utilizamos as redes de coautoria gr-qc, cond-mat, astro-ph, hep-ph e hep-th do *ArXiv* com informações de publicações entre os anos 1992 e 1998.

Produzimos os grafos G_t e $G_{t'}$ com $t < t'$ que representam a rede, para valores $t = 1997$ e $t' = 1998$.

Observou-se que existem 394 arestas no *dataset* gr-qc, 2.398 arestas no *dataset* cond-mat, 10.123 arestas no *dataset* astro-ph, 2.306 arestas no *dataset* hep-ph e 1.215 arestas no *dataset* hep-th. As arestas novas são do grafo G_{1998} entre pares de nós não adjacentes em G_{1997} para os cinco *datasets* em questão. Neste experimento, devido ao tamanho dos grafos e à variação de valores dos *scores* obtidos, precisou-se considerar valores de Top-k convenientes.

Conforme foi descrito na Seção 5.2.1, foram considerados 2 experimentos, um para comparar a métrica tradicional e aos pares versão 'ou' (Experimento 2.1) e o segundo para comparar com a métrica aos pares versão 'e' (Experimento 2.2).

Em todos os casos dos Experimentos 2.1 e 2.2, os valores de k para calcular o Top-k foram escolhidos após realizar testes para obter valores de VP (verdadeiros positivos) diferentes de zero.

5.3.2 Resultados obtidos do experimento 2

Todos os valores das medidas de qualidade de classificação dos *links* preditos obtidos neste experimento para o dataset gr-qc são apresentados na Tabela 10 do Experimento 2.1. A Tabela 10 apresenta resultados para $G_t = G_{1997}$ e $G_{t'} = G_{1998}$ para Vizinhos Comuns tradicional e aos pares com valor Top-7 para o *dataset* gr-qc do Experimento 2.1. O valor $k = 7$ foi escolhido porque a métrica tradicional Vizinhos Comuns apresentou unicamente 9 valores de *scores* diferentes.

Para as métricas Similaridade de Jaccard e Adamic-Adar tradicional e aos pares, foi utilizado o Top-25. O valor $k = 25$ foi escolhido porque, no caso dessas duas métricas, os valores de *scores* diferentes obtidos foram em maior quantidade, comparados à métrica Vizinhos Comuns. Para Ligação Preferencial tradicional e aos pares foi utilizado o Top-100. Foi escolhido o Top-100 para que tivesse ao menos uma aresta acima do Top, ou seja, pelo menos um Verdadeiro Positivo (VP), tornando, assim, as medidas de Precisão, F-1 e Revocação diferentes de zero.

Assim, pode-se concluir que, para Vizinhos Comuns e Jaccard, os resultados obtidos pelo método aos pares foram melhores, exceto na Revocação. Para Ligação Preferencial, o método aos pares teve melhor resultado em todos os casos. E para Adamic-Adar, os resultados obtidos pelo método aos pares foram melhores, exceto na Acurácia. Além disso, foi possível concluir que o preditor randômico não teve bom resultado para as quatro métricas na Precisão.

Tabela 10 – Resultados para G_{1997}, G_{1998} no *dataset* gr-qc do Experimento 2.1

Top-7				
Dataset	Pred Rand	Medidas de	Métrica	
gr-qc	0,00011494	qualidade	VC	VC^*
		Precisão	0,008254717	**0,026845638**
		Acurácia	0,999155188	**0,9998427**
		F-1	0,014295439	**0,014625229**
		Revocação	**0,053299492**	0,010050251
Top-25				
			JS	JS^*
		Precisão	0,005237712	**0,006445672**
		Acurácia	0,999168607	**0,999570**
		F-1	0,009040334	**0,009414929**
		Revocação	**0,032994924**	0,017456359
Top-100				
			LP	LP^*
		Precisão	0,001038422	**0,003809524**
		Acurácia	0,999604724	**0,999732**
		F-1	0,001473839	**0,004343105**
		Revocação	0,002538071	**0,005050505**
Top-25				
			AA	AA^*
		Precisão	0,018691589	**0,027210884**
		Acurácia	**0,999855017**	0,9998433
		F-1	0,007984032	**0,014678899**
		Revocação	0,005076142	**0,010050251**

A Tabela 11 apresenta resultados para $G_t = G_{1997}$ e $G_{t'} = G_{1998}$ para Vizinhos Comuns tradicional e aos pares com valor Top-7 para o *dataset* cond-mat do Experimento 2.1. O valor $k = 7$ foi escolhido porque a métrica tradicional Vizinhos Comuns apresentou unicamente 10 valores de *scores* diferentes.

Para a métrica Similaridade de Jaccard tradicional e aos pares foi utilizado o Top-25. O valor $k = 25$ foi escolhido porque, para a métrica Similaridade de Jaccard, os valores de *scores* diferentes obtidos foram em maior quantidade do que Vizinhos Comuns.

Para a métrica Ligação Preferencial tradicional e aos pares foi escolhido o Top-201 e para a métrica Adamic-Adar foi escolhido o Top-100 para que tivessem ao menos uma aresta que apareceu em G_{1998}, ou seja, acima do Top, tornando, assim, as medidas de Precisão, F-1 e Revocação diferentes de zero.

Tabela 11 – Resultados para G_{1997}, G_{1998} no *dataset* cond-mat do Experimento 2.1

Top-7				
Dataset	Pred Rand	Medidas de	Métrica	
cond-mat	0,00006877	qualidade	VC	VC^*
		Precisão	0,015745999	**0,029850746**
		Acurácia	0,99982363	**0,999929366**
		F-1	**0,019451531**	0,001621403
		Revocação	**0,025437865**	0,000833333
Top-25				
			JS	JS^*
		Precisão	0,009633911	**0,010736196**
		Acurácia	0,999858245	**0,99991273**
		F-1	**0,01001402**	0,004579653
		Revocação	**0,010425354**	0,002910603
Top-201				
			LP	LP^*
		Precisão	**0,002083333**	0,001203369
		Acurácia	0,999890106	**0,9999074**
		F-1	**0,001563314**	0,000619195
		Revocação	**0,001251043**	0,00041684
Top-100				
			AA	AA^*
		Precisão	**0,026315789**	0,016251354
		Acurácia	**0,999922971**	0,99990519
		F-1	0,00592154	**0,008992806**
		Revocação	0,003336113	**0,006216328**

Todos os valores das medidas de qualidade de classificação dos *links* preditos obtidos neste experimento para o dataset cond-mat são apresentados na Tabela 11 do Experimento 2.1. Assim, pode-se concluir que, para Vizinhos Comuns, Similaridade de Jaccard e Adamic-Adar, os resultados obtidos foram balanceados pelos dois métodos. Para Ligação Preferencial, o método tradicional teve melhor resultado em todos os casos, exceto na Acurácia. Além disso, foi possível concluir que o preditor randômico não teve bom resultado para as quatro métricas na Precisão.

A Tabela 12 apresenta resultados para $G_t = G_{1997}$ e $G_{t'} = G_{1998}$ para Vizinhos Comuns tradicional e aos pares com valor Top-7 para o *dataset* astro-ph do Experimento 2.1. O valor $k = 7$ foi escolhido porque a métrica tradicional Vizinhos Comuns apresentou 34 valores de *scores* diferentes.

Para as métricas Similaridade de Jaccard, Ligação Preferencial e Adamic-Adar tradicional e aos pares foi escolhido o Top-25. O valor $k = 25$ foi escolhido porque os valores de *scores* diferentes obtidos foram em maior quantidade comparados a Vizinhos Comuns.

Tabela 12 – Resultados para G_{1997}, G_{1998} no *dataset* astro-ph do Experimento 2.1

Top-7				
Dataset	Pred Rand	Medidas de	Métrica	
astro-ph	0,00031114	qualidade	VC	VC^*
		Precisão	0,060465116	**0,085470085**
		Acurácia	0,999683049	**0,99968557**
		F-1	**0,002514993**	0,00195122
		Revocação	**0,001284204**	0,000986875
Top-25				
			JS	JS^*
		Precisão	**0,054054054**	0,006430868
		Acurácia	**0,999682773**	0,99967936
		F-1	**0,002319961**	0,000383289
		Revocação	**0,001185419**	0,000197531
Top-25				
			LP	LP^*
		Precisão	0,037037037	0,037037037
		Acurácia	**0,99968809**	0,99968806
		F-1	**0,000197044**	0,000197025
		Revocação	**0,0000987849**	0,0000987752
Top-25				
			AA	AA^*
		Precisão	0,060913706	**0,114503817**
		Acurácia	0,999683541	**0,99968529**
		F-1	0,002325581	**0,002921414**
		Revocação	0,001185419	**0,001479582**

Todos os valores das medidas de qualidade de classificação dos *links* preditos obtidos neste experimento para o *dataset* astro-ph são apresentados na Tabela 12 do Experimento 2.1. Assim, pode-se concluir que, para Vizinhos Comuns, os resultados obtidos foram balanceados pelos dois métodos.

Para Similaridade de Jaccard, o método tradicional teve melhor resultado em todos os casos. Na métrica Ligação Preferencial, os resultados foram iguais ou melhores para o método tradicional. Além disso, foi possível concluir que o preditor randômico não teve bom resultado para as quatro métricas na Precisão.

A Tabela 13 apresenta resultados para $G_t = G_{1997}$ e $G_{t'} = G_{1998}$ para o *dataset* hep-ph do Experimento 2.1.

Para as métricas Vizinhos Comuns e Similaridade de Jaccard tradicional e aos pares foi escolhido o Top-25. O valor $k = 25$ foi escolhido porque, no caso das métricas Vizinhos Comuns e Similaridade de Jaccard, com esse Top já foi suficiente para ter ao menos uma aresta como Verdadeiro Positivo (VP), ou seja, uma aresta acima do Top.

Para a métrica Ligação Preferencial tradicional e aos pares foi escolhido o Top-300 e para a métrica Adamic-Adar foi escolhido o Top-100 para que tivessem ao menos uma aresta acima do Top, tornando, assim, as medidas de Precisão, F-1 e Revocação diferentes de zero. Ainda assim, em Ligação Preferencial aos pares, os resultados deram zero em Precisão, F-1 e Revocação pois não foram encontradas arestas que apareceram em G_{1998} acima do Top-300.

Tabela 13 – Resultados para G_{1997}, G_{1998} no *dataset* hep-ph do Experimento 2.1

Top-25				
Dataset	Pred Rand	Medidas de	Métrica	
hep-ph	0,00010109	qualidade	VC	VC^*
		Precisão	**0,001251043**	0,000705716
		Acurácia	0,999689205	**0,9998368**
		F-1	**0,001689665**	0,000537057
		Revocação	**0,002601908**	0,000433463
Top-25				
			JS	JS^*
		Precisão	**0,013207547**	0,002331002
		Acurácia	**0,999876296**	0,99986139
		F-1	**0,00493653**	0,001263424
		Revocação	**0,003035559**	0,000866551
Top-300				
			LP	LP^*
		Precisão	**0,001686341**	0
		Acurácia	**0,999873008**	0,99985775
		F-1	**0,000689893**	0
		Revocação	**0,000433651**	0
Top-100				
			AA	AA^*
		Precisão	**0,004504505**	0,00101626
		Acurácia	**0,999879627**	0,9998558
		F-1	**0,001454545**	0,000607718
		Revocação	**0,000867303**	0,000433463

Todos os valores das medidas de qualidade de classificação dos *links* preditos obtidos neste experimento para o dataset hep-ph são apresentados na Tabela 13 do Experimento 2.1. Assim, pode-se concluir que, para todas as quatro métricas, os resultados obtidos foram melhores para o método tradicional, exceto em Vizinhos Comuns na Acurácia. Além disso, foi possível concluir que o preditor randômico não teve bom resultado para as quatro métricas, exceto em Ligação Preferencial aos pares em Precisão, onde os resultados deram zero.

A Tabela 14 apresenta resultados para $G_t = G_{1997}$ e $G_{t'} = G_{1998}$ para Vizinhos Comuns tradicional e aos pares com valor Top-7 para o *dataset* hep-th do Experimento

2.1. O valor $k = 7$ foi escolhido porque a métrica tradicional Vizinhos Comuns apresentou unicamente 10 valores de *scores* diferentes.

Para a métrica Similaridade de Jaccard tradicional e aos pares, foi escolhido o Top-25. O valor $k = 25$ foi escolhido porque os valores de *scores* diferentes obtidos foram em maior quantidade em comparação a Vizinhos Comuns.

Para as métricas Ligação Preferencial e para Adamic-Adar tradicional e aos pares foi escolhido o Top-100 para que tivessem ao menos uma aresta que apareceu em G_{1998} acima do Top, tornando, assim, as medidas de Precisão, F-1 e Revocação diferentes de zero.

Tabela 14 – Resultados para G_{1997}, G_{1998} no *dataset* hep-th do Experimento 2.1

Top-7				
Dataset	Pred Rand	Medidas de	Métrica	
hep-th	0,0000625	qualidade	VC	VC^*
		Precisão	**0,021237864**	0,020484171
		Acurácia	0,99985633	**0,99991044**
		F-1	**0,024449878**	0,012478729
		Revocação	**0,028806584**	0,008972268
Top-25				
			JS	JS^*
		Precisão	**0,007527181**	0,005427408
		Acurácia	0,999755766	**0,999899**
		F-1	**0,011245314**	0,00408998
		Revocação	**0,022222222**	0,003281378
Top-100				
			LP	LP^*
		Precisão	**0,005263158**	0,005235602
		Acurácia	0,999898819	**0,99991795**
		F-1	**0,004050633**	0,002501563
		Revocação	**0,003292181**	0,001643385
Top-100				
			AA	AA^*
		Precisão	**0,026785714**	0,013386881
		Acurácia	**0,999932049**	0,99989959
		F-1	0,004521477	**0,010141988**
		Revocação	0,002469136	**0,008163265**

Todos os valores das medidas de qualidade de classificação dos *links* preditos obtidos neste experimento para o *dataset* hep-th são apresentados na Tabela 14 do Experimento 2.1. Assim, pode-se concluir que, para as métricas Vizinhos Comuns, Similaridade de Jaccard e Ligação Prefencial, os resultados obtidos foram melhores para o método tradicional, exceto na Acurácia. Para a métrica Adamic-Adar, os resultados obtidos foram balanceados para

os dois métodos. Além disso, foi possível concluir que o preditor randômico não teve bom resultado para as quatro métricas na Precisão.

A Tabela 15 apresenta resultados para $G_t = G_{1997}$ e $G_{t'} = G_{1998}$ para Vizinhos Comuns tradicional e aos pares com valor Top-7 para o *dataset* gr-qc do Experimento 2.2. O valor $k = 7$ foi escolhido porque a métrica tradicional Vizinhos Comuns apresentou unicamente 9 valores de *scores* diferentes.

Para a métrica Similaridade de Jaccard tradicional e aos pares, foi utilizado o Top-25. O valor $k = 25$ foi escolhido porque os valores de *scores* diferentes obtidos foram em maior quantidade, comparados à métrica Vizinhos Comuns.

Para a métrica Ligação Preferencial tradicional e aos pares foi escolhido o Top-100 e para a métrica Adamic-Adar foi escolhido Top-45 para que tivessem ao menos uma aresta que apareceu em G_{1998} acima do Top, tornando, assim, as medidas de Precisão, F-1 e Revocação diferentes de zero.

Tabela 15 – Resultados para G_{1997}, G_{1998} no *dataset* gr-qc do Experimento 2.2

Top-7				
Dataset	Pred Rand	Medidas de	Métrica	
gr-qc	0,00011494	qualidade	VC	VC^{**}
		Precisão	**0,008254717**	0,007675906
		Acurácia	0,999155188	**0,99920**
		F-1	**0,014295439**	0,0130
		Revocação	**0,053299492**	0,043
Top-25				
			JS	JS^{**}
		Precisão	0,005237712	**0,007567219**
		Acurácia	**0,999168607**	0,9980
		F-1	0,009040334	**0,01413**
		Revocação	0,032994924	**0,106**
Top-100				
			LP	LP^{**}
		Precisão	**0,001038422**	0,000563698
		Acurácia	**0,999604724**	0,99936
		F-1	**0,001473839**	0,000922084
		Revocação	**0,002538071**	0,002531646
Top-45				
			AA	AA^{**}
		Precisão	**0,014492754**	0,009615385
		Acurácia	0,999845974	**0,9998550**
		F-1	**0,007518797**	0,004008016
		Revocação	**0,005076142**	0,002531646

Todos os valores das medidas de qualidade de classificação dos *links* preditos obtidos

neste experimento para o *dataset* gr-qc são apresentados na Tabela 15 do Experimento 2.2. Assim, pode-se concluir que, para Vizinhos Comuns, os resultados obtidos pelo método tradicional foi melhor, exceto na Acurácia. Para Similaridade de Jaccard, o resultado foi melhor aos pares em todas as medidas de qualidade, exceto na Acurácia. Para Ligação Preferencial, o método tradicional teve melhor resultado em todos os casos. E para Adamic-Adar, o método tradicional teve melhor resultado, exceto na Acurácia. Além disso, foi possível concluir que o preditor randômico não teve bom resultado para as quatro métricas na Precisão.

A Tabela 16 apresenta resultados para $G_t = G_{1997}$ e $G_{t'} = G_{1998}$ para Vizinhos Comuns tradicional e aos pares com valor Top-7 para o *dataset* cond-mat do Experimento 2.2. O valor $k = 7$ foi escolhido porque a métrica tradicional Vizinhos Comuns apresentou unicamente 10 valores de *scores* diferentes.

Para a métrica Similaridade de Jaccard tradicional e aos pares, foi escolhido o Top-25. O valor $k = 25$ foi escolhido porque os valores de *scores* diferentes obtidos foram em maior quantidade, comparados a Vizinhos Comuns.

Para a métrica Ligação Preferencial tradicional e aos pares foi escolhido o Top-201 e para a métrica Adamic-Adar foi escolhido Top-100 para que tivessem ao menos uma aresta do conjunto de teste acima do Top, tornando, assim, as medidas de Precisão, F-1 e Revocação diferentes de zero.

Todos os valores das medidas de qualidade de classificação dos *links* preditos obtidos neste experimento para o *dataset* cond-mat são apresentados na Tabela 16 do Experimento 2.2. Assim, pode-se concluir que, para Vizinhos Comuns e Adamic-Adar, os resultados obtidos foram melhores pelo método tradicional, exceto na Acurácia. Para Jaccard, o método aos pares teve melhor resultado em todos os casos, exceto na Acurácia. E na métrica Ligação Preferencial, os resultados foram balanceados para os dois métodos. Além disso, foi possível concluir que o preditor randômico não teve bom resultado para as quatro métricas na Precisão.

Tabela 16 – Resultados para G_{1997}, G_{1998} no *dataset* cond-mat do Experimento 2.2

Top-7				
Dataset	Pred Rand	Medidas de	Métrica	
cond-mat	0,00006877	qualidade	VC	VC^{**}
		Precisão	**0,015745999**	0,01541976
		Acurácia	0,99982363	**0,999832**
		F-1	**0,019451531**	0,0181
		Revocação	**0,025437865**	0,0220
Top-25				
			JS	JS^{**}
		Precisão	0,009633911	**0,010610348**
		Acurácia	**0,999858245**	0,999650
		F-1	0,01001402	**0,01693**
		Revocação	0,010425354	**0,0419**
Top-201				
			LP	LP^{**}
		Precisão	**0,002083333**	0,001877582
		Acurácia	**0,999890106**	0,9998550
		F-1	0,001563314	**0,001973944**
		Revocação	0,001251043	**0,002080732**
Top-100				
			AA	AA^{**}
		Precisão	**0,026315789**	0,01875
		Acurácia	0,999922971	**0,999926728**
		F-1	**0,00592154**	0,002342835
		Revocação	**0,003336113**	0,001249479

A Tabela 17 apresenta resultados para $G_t = G_{1997}$ e $G_{t'} = G_{1998}$ para Vizinhos Comuns tradicional e aos pares com valor Top-7 para o *dataset* astro-ph do Experimento 2.2. O valor $k = 7$ foi escolhido porque a métrica tradicional Vizinhos Comuns apresentou 34 valores de *scores* diferentes.

Para a métrica Similaridade de Jaccard tradicional e aos pares, foi escolhido o Top-25. O valor $k = 25$ foi escolhido porque os valores de *scores* diferentes obtidos foram em maior quantidade em relação a Vizinhos Comuns.

Para a métrica Ligação Preferencial tradicional e aos pares foi escolhido o Top-35 e para a métrica Adamic-Adar foi escolhido Top-25 para que tivessem ao menos uma aresta do conjunto de teste acima do Top, tornando, assim, as medidas de Precisão, F-1 e Revocação diferentes de zero.

Tabela 17 – Resultados para G_{1997}, G_{1998} no *dataset* astro-ph do Experimento 2.2

Top-7				
Dataset	Pred Rand	Medidas de	Métrica	
astro-ph	0,00031114	qualidade	VC	VC^{**}
		Precisão	0,060465116	**0,080536913**
		Acurácia	0,999683049	**0,9996846**
		F-1	**0,002514993**	0,002333
		Revocação	**0,001284204**	0,001184
Top-25				
			JS	JS^{**}
		Precisão	**0,054054054**	0,013944223
		Acurácia	**0,999682773**	0,9996432
		F-1	0,002319961	**0,003605**
		Revocação	0,001185419	**0,002070**
Top-35				
			LP	LP^{**}
		Precisão	**0,052631579**	0,032258065
		Acurácia	**0,999687813**	0,99968701
		F-1	**0,000393662**	0,000392657
		Revocação	**0,00019757**	0,000197531
Top-25				
			AA	AA^{**}
		Precisão	0,060913706	**0,454545455**
		Acurácia	0,999683541	**0,999688**
		F-1	**0,002325581**	0,001969
		Revocação	**0,001185419**	0,000986

Todos os valores das medidas de qualidade de classificação dos *links* preditos obtidos neste experimento para o *dataset* astro-ph são apresentados na Tabela 17 do Experimento 2.2. Assim, pode-se concluir que, para Vizinhos Comuns, Similaridade de Jaccard e Adamic-Adar os resultados obtidos foram balanceados pelos dois métodos. Para Similaridade de Jaccard, o método tradicional teve melhor resultado em todos os casos. Na métrica Ligação Preferencial, os resultados foram melhores para o método tradicional. Além disso, foi possível concluir que o preditor randômico não teve bom resultado para as quatro métricas na Precisão.

A Tabela 18 apresenta resultados para $G_t = G_{1997}$ e $G_{t'} = G_{1998}$ para o *dataset* hep-ph do Experimento 2.2.

Para as métricas Vizinhos Comuns e Similaridade de Jaccard tradicional e aos pares, foi escolhido o Top-25. O valor $k = 25$ foi escolhido porque já foi suficiente para ter ao menos uma aresta do conjunto de teste acima do Top, tornando, assim, as medidas de Precisão, F-1 e Revocação diferentes de zero.

Para a métrica Ligação Preferencial tradicional e aos pares foi escolhido o Top-300 e para a métrica Adamic-Adar foi escolhido Top-100 para que tivessem ao menos uma aresta que apareceu em G_{1998} acima do Top, tornando, assim, as medidas de Precisão, F-1 e Revocação diferentes de zero. Ainda assim, em Ligação Preferencial e Adamic-Adar aos pares, Precisão, F-1 e Revocação deram zero pois não tiveram arestas acima do Top selecionado.

Tabela 18 – Resultados para G_{1997}, G_{1998} no *dataset* hep-ph do Experimento 2.2

Top-25				
Dataset	Pred Rand	Medidas de	Métrica	
hep-ph	0,00010109	qualidade	VC	VC^{**}
		Precisão	0,001251043	**0,00133936**
		Acurácia	**0,999689205**	0,999637
		F-1	0,001689665	**0,001930735**
		Revocação	0,002601908	**0,003457217**
Top-25				
			JS	JS^{**}
		Precisão	**0,013207547**	0,011811024
		Acurácia	**0,999876296**	0,999799
		F-1	0,00493653	**0,011690842**
		Revocação	0,003035559	**0,011573082**
Top-300				
			LP	LP^{**}
		Precisão	**0,001686341**	0
		Acurácia	0,999873008	**0,99988440**
		F-1	**0,000689893**	0
		Revocação	**0,000433651**	0
Top-100				
			AA	AA^{**}
		Precisão	**0,004504505**	0
		Acurácia	0,999879627	**0,9998955**
		F-1	**0,001454545**	0
		Revocação	**0,000867303**	0

Todos os valores das medidas de qualidade de classificação dos *links* preditos obtidos neste experimento para o *dataset* hep-ph são apresentados na Tabela 18 do Experimento 2.2. Assim, pode-se concluir que, para Vizinhos Comuns, os resultados obtidos foram melhores para o método aos pares, exceto na Acurácia. Para Similaridade de Jaccard, os resultados obtidos foram balanceados para os dois métodos. Nas métricas Ligação Preferencial e Adamic-Adar, os resultados foram melhores para o método tradicional, exceto na Acurácia. Além disso, foi possível concluir que o preditor randômico não teve bom resultado para as quatro métricas, exceto em Ligação Preferencial e Adamic-Adar aos pares em Precisão, pois os resultados deram zero.

A Tabela 19 apresenta resultados para $G_t = G_{1997}$ e $G_{t'} = G_{1998}$ para Vizinhos Comuns tradicional e aos pares com valor Top-5 para o *dataset* hep-th do Experimento 2.2. O valor $k = 5$ foi escolhido porque a métrica tradicional Vizinhos Comuns apresentou unicamente 10 valores de *scores* diferentes.

Para a métrica Similaridade de Jaccard tradicional e aos pares foi escolhido o Top-25. O valor $k = 25$ foi escolhido porque os valores de *scores* diferentes obtidos foram em maior quantidade em relação a Vizinhos Comuns.

Para as métricas Ligação Preferencial e Adamic-Adar, tradicional e aos pares, foi escolhido o Top-100 para que tivessem ao menos uma aresta do conjunto de teste acima do Top, tornando, assim, as medidas de Precisão, F-1 e Revocação diferentes de zero.

Tabela 19 – Resultados para G_{1997}, G_{1998} no *dataset* hep-th do Experimento 2.2

Top-5				
Dataset	Pred Rand	Medidas de	Métrica	
hep-th	0,0000625	qualidade	VC	VC^{**}
		Precisão	**0,035714286**	0,010353095
		Acurácia	**0,999932152**	0,999470
		F-1	0,006028636	**0,0181**
		Revocação	0,003292181	**0,072**
Top-25				
			JS	JS^{**}
		Precisão	**0,007527181**	0,007097038
		Acurácia	**0,999755766**	0,999189
		F-1	0,011245314	**0,01302**
		Revocação	0,022222222	**0,07**
Top-100				
			LP	LP^{**}
		Precisão	**0,005263158**	0,000904159
		Acurácia	**0,999898819**	0,99976
		F-1	**0,004050633**	0,001322751
		Revocação	**0,003292181**	0,002463054
Top-100				
			AA	AA^{**}
		Precisão	**0,026785714**	0,008368201
		Acurácia	**0,999932049**	0,99992531
		F-1	**0,004521477**	0,002747253
		Revocação	**0,002469136**	0,001643385

Todos os valores das medidas de qualidade de classificação dos *links* preditos obtidos neste experimento para o *dataset* hep-th são apresentados na Tabela 19 do Experimento 2.2. Assim, pode-se concluir que, para as métricas Vizinhos Comuns e Similaridade de Jaccard, os resultados obtidos foram balanceados para os dois métodos. Para as métricas Ligação Preferencial e Adamic-Adar, os resultados obtidos foram melhores para o método

tradicional. Além disso, foi possível concluir que o preditor randômico não teve bom resultado para as quatro métricas na Precisão.

Para os dois experimentos, observou-se que a Acurácia deu valores altos para todas as métricas, devido à quantidade alta de Verdadeiros Negativos.

6 JUSTIFICATIVA E COMPARAÇÃO

Neste capítulo é feita uma síntese dos resultados obtidos no capítulo anterior. Primeiro, é feito um resumo do resultado obtido na comparação de todas as métricas na versão tradicional e aos pares (‘ou’) e com aos pares (‘e’).

As Tabelas 20 e 21 apresentam um resumo dos resultados obtidos que permite observar os casos em que a versão aos pares superou o resultado da versão tradicional. A notação ‘*’ nas tabelas significa que a versão ‘aos pares’ teve melhor desempenho que a tradicional.

Os resultados da comparação da rede Lattes estão representados na Tabela 20. A primeira coluna representa a versão tradicional x aos pares ‘ou’ e a segunda coluna mostra a versão tradicional x aos pares ‘e’.

Tabela 20 – Tabela comparativa das versões aos pares ‘ou’ e ‘e’ da rede Lattes

G_{2011}-G_{2014}									
Aos pares ‘ou’ melhor que tradicional					**Aos pares ‘e’ melhor que tradicional**				
métrica	**Precisão**	**Acurácia**	**F-1**	**Revoc**	**métrica**	**Precisão**	**Acurácia**	**F-1**	**Revoc**
VC*	*	*	*		VC**				*
JS*		*			JS**	*		*	*
LP*					LP**				
AA*			*	*	AA**				

Na rede Lattes, a métrica Vizinhos Comuns teve melhor desempenho na versão aos pares em detrimento da métrica tradicional. A métrica ‘ou’ teve melhor resultado em Precisão, Acurácia e F-1. A métrica ‘e’ teve melhor resultado em Revocação. A métrica Similaridade de Jaccard teve melhor desempenho na versão aos pares em Acurácia na métrica ‘ou’ e em Precisão, F-1 e Revocação na métrica ‘e’. Para a métrica Ligação Preferencial, a versão tradicional teve melhor resultado nas quatro medidas de qualidade (Precisão, Acurácia, F-1 e Revocação). Já a métrica Adamic-Adar, a versão aos pares teve melhor desempenho em F-1 e Revocação na versão ‘ou’. Na versão ‘e’, a métrica tradicional teve melhor resultado.

Os resultados da comparação para os cinco *datasets* do *ArXiv* (gr-qc, cond-mat, astro-ph, hep-ph e hep-th) estão representados na Tabela 21. A primeira coluna representa a versão tradicional x aos pares ‘ou’ e a segunda coluna mostra a versão tradicional x aos pares ‘e’.

Tabela 21 – Tabela comparativa das versões aos pares 'ou' e 'e' das redes do *ArXiv*.

Gr-qc									
Aos pares 'ou' melhor que tradicional					**Aos pares 'e' melhor que tradicional**				
métrica	**Precisão**	**Acurácia**	**F-1**	**Revoc**	**métrica**	**Precisão**	**Acurácia**	**F-1**	**Revoc**
VC*	*	*	*		VC**		*		
JS*	*	*	*		JS**	*		*	*
LP*	*	*	*	*	LP**				
AA*	*		*	*	AA**		*		
Cond-mat									
Aos pares 'ou' melhor que tradicional					**Aos pares 'e' melhor que tradicional**				
métrica	**Precisão**	**Acurácia**	**F-1**	**Revoc**	**métrica**	**Precisão**	**Acurácia**	**F-1**	**Revoc**
VC*	*	*			VC**		*		
JS*	*	*			JS**	*		*	*
LP*		*			LP**			*	*
AA*			*	*	AA**		*		
Astro-ph									
Aos pares 'ou' melhor que tradicional					**Aos pares 'e' melhor que tradicional**				
métrica	**Precisão**	**Acurácia**	**F-1**	**Revoc**	**métrica**	**Precisão**	**Acurácia**	**F-1**	**Revoc**
VC*	*	*			VC**	*	*		
JS*					JS**			*	*
LP*					LP**				
AA*	*	*	*	*	AA**	*	*		
Hep-ph									
Aos pares 'ou' melhor que tradicional					**Aos pares 'e' melhor que tradicional**				
métrica	**Precisão**	**Acurácia**	**F-1**	**Revoc**	**métrica**	**Precisão**	**Acurácia**	**F-1**	**Revoc**
VC*		*			VC**	*		*	*
JS*					JS**			*	*
LP*					LP**		*		
AA*					AA**		*		
Hep-th									
Aos pares 'ou' melhor que tradicional					**Aos pares 'e' melhor que tradicional**				
métrica	**Precisão**	**Acurácia**	**F-1**	**Revoc**	**métrica**	**Precisão**	**Acurácia**	**F-1**	**Revoc**
VC*		*			VC**			*	*
JS*		*			JS**			*	*
LP*		*			LP**				
AA*			*	*	AA**				

Nas redes do *ArXiv*, a versão aos pares 'ou' apresentou melhores resultados que a versão tradicional nos seguintes casos:

- para Acurácia:
  - em todas as 5 redes para a métrica VC*,
  - em 3 das redes para a métrica JS* (gr-qc, cond-mat e hep-th),
  - em 3 das redes para a métrica LP* (gr-qc, cond-mat e hep-th)
- para Precisão:
  - em 3 das redes para a métrica VC* (gr-qc, cond-mat e astro-ph),
  - em 2 das redes para a métrica JS* (gr-qc e cond-mat)

- para F-1 e Revocação:
  - em 4 das redes para a métrica AA* (gr-qc, cond-mat, astro-ph e hep-th),

A versão aos pares 'e' apresentou melhores resultados que a versão aos pares 'ou' nos seguintes casos:

- para F-1 e Revocação:
  - em todas as 5 redes para a métrica JS**,
  - em 2 das redes para a métrica VC** (hep-ph e hep-th)

Observou-se que a versão 'e' é mais restritiva que a versão 'ou', pois o número de elementos na interseção é menor ou igual ao número de elementos na união de dois conjuntos. Dessa forma, podemos atribuir a esse fato que a versão 'ou' teve melhores resultados que a versão 'e'.

Além disso, constatou-se que o valor da transitividade é alto para os *datasets* gr-qc e hep-ph. Entretanto, somente no *dataset* gr-qc a métrica aos pares teve melhor resultado na versão 'ou'. Dessa forma, não podemos afirmar que a transitividade é determinante para obter melhores resultados de alguma das duas versões das métricas.

7 CONCLUSÃO

Devido à crescente demanda por utilização de redes sociais, muitos pesquisadores concentraram seus estudos na análise da evolução das redes ao longo do tempo. Por este motivo, novas técnicas de predição de *links*, problemas e aplicações estão surgindo rapidamente com o passar do tempo (WANG et al., 2015). Recentemente, (NASSAR; BENSON; GLEICH, 2019a) (NASSAR; BENSON; GLEICH, 2020) propuseram uma nova versão de métrica topológica que foi denominada predição de *links* aos pares. Nesta nova abordagem, considera-se uma aresta na rede e tenta descobrir quais são os nós prováveis de se conectar no futuro com essa determinada aresta.

O presente trabalho se propôs a fazer uma análise comparativa das métricas topológicas locais tradicional e aos pares, utilizando os métodos 'ou' e 'e'. Utilizando a abordagem não supervisionada de predição de *links*, o presente trabalho avaliou quatro métricas de similaridade em uma rede da Plataforma Lattes e em cinco redes de coautoria do *ArXiv*. A partir dos resultados obtidos, pode-se concluir que as duas abordagens para as quatro métricas consideradas, Vizinhos Comuns, Jaccard, Ligação Preferencial e Adamic-Adar, apresentam comportamentos parecidos, com uma pequena vantagem para a versão aos pares. Pela observação dos resultados preliminares, pode-se considerar que a versão aos pares, que foi introduzida recentemente, também pode ser usada para resolver a abordagem topológica do problema de predição de *links*. Concluiu-se que a versão 'e' é mais restritiva que a versão 'ou', pois o número de elementos na interseção é menor ou igual ao número de elementos na união de dois conjuntos. Logo, se atribui a esse fato que a versão 'ou' teve melhores resultados que a versão 'e'. A pequena vantagem da métrica aos pares na versão 'ou' parece estar relacionada com a definição da métrica. Além disso, constatou-se, pelos experimentos, que a transitividade não parece influenciar diretamente nos resultados.

Como possíveis trabalhos futuros, propomos realizar outros experimentos utilizando períodos de tempo diferentes para os conjuntos de treino e teste (G_t e $G_{t'}$). Além disso, propomos estender os experimentos realizados considerando outros valores de Top-k com o objetivo de avaliar melhor alguns dos *datasets* do *ArXiv* utilizados. Novos experimentos podem ser realizados utilizando outras redes sociais. Também como trabalho futuro, poderá ser utilizada a abordagem supervisionada, usando métricas que proveem atributos para classificação, como a informação contextual, adicionando-se a informação temporal, combinando as informações contextual e temporal.

Como generalização do trabalho realizado nesta dissertação, podem ser consideradas estruturas mais complexas para definir vizinhanças (K_k para $k \geq 4$), de forma equivalente a que foi realizada por (NASSAR; BENSON; GLEICH, 2019a) (NASSAR; BENSON; GLEICH, 2020) para triângulos (K_3) ou por (LIBEN-NOWELL; KLEINBERG, 2007)

para arestas (K_2).

REFERÊNCIAS

ADAMIC, L. A.; ADAR, E. Friends and neighbors on the web. *Social Networks*, v. 25, 2003. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0378873303000091>.

ANDERSEN, R.; CHUNG, F.; LANG, K. Local graph partitioning using pagerank vectors. In: *47th Annual IEEE Symposium*. Foundations of Computer Science (FOCS'06): IEEE, 2006. p. 475–486.

BACKSTROM, L.; LESKOVEC, J. Supervised random walks: Predicting and recommending links in social networks. In: *Proceedings of the fourth ACM international conference on Web search and data mining*. Hong Kong, China: ACM, 2011. p. 635–644.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. *science*, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999.

BARBOSA, D. A. de B. L.; AVELINO, L. B.; SOUZA, R. F. de; OLIVEIRA, C. C. G. F. de; JUSTEL, C. Medidas de centralidade e detecção de comunidades em rede de co-autoria. In: *Anais do XLIII Simpósio Brasileiro de Pesquisa Operacional*. Ubatuba: Simpósio Brasileiro de Pesquisa Operacional, 2011. p. pp. 2574–2583.

BENCHETTARA, N.; KANAWATI, R.; ROUVEIROL, C. Supervised machine learning applied to link prediction in bipartite social networks. In: *2010 International Conference on Advances in Social Networks Analysis and Mining*. NW Washington, DC, United States: IEEE Computer Society, 2010. p. 326–330.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, Elsevier, v. 30, n. 1-7, p. 107–117, 1998.

CAVALCANTE, A. A. B.; JUSTEL, C. M.; GOLDSCHMIDT, R. R. Link prediction in social networks: an edge creation history retrieval based method that combines topological and contextual data. In: *Lecture Notes in Computer Science 12320*. [S.l.]: Springer International Publishing, 2020. p. 382–396.

CLAUSET, A.; MOORE, C.; NEWMAN, M. E. J. Hierarchical structure and the prediction of missing links in networks. *Nature*, Nature Publishing Group, v. 453, n. 7191, p. 98–101, 2008.

DAVIS, D.; LICHTENWALTER, R.; CHAWLA, N. V. Supervised methods for multi-relational link prediction. *Social network analysis and mining*, Springer, v. 3, n. 2, p. 127–141, 2013.

EASLEY, D.; KLEINBERG, J. Networks, crowds, and markets: Reasoning about a highly connected world. *ISBN 0521195330, 9780521195331*, 2010.

FLORENTINO, E. da S.; GOLDSCHMIDT, R. R. *Utilizando o Histórico da Evolução de Redes Complexas na Tarefa de Predição de Link*. 120 p. Mestrado em Sistemas e Computação — Instituto Militar de Engenharia, Rio de Janeiro, 2017.

FORTUNATO, S. Community detection in graphs. *Physics Reports*, v. 486, n. 3, p. 75 – 174, 2010. ISSN 0370-1573. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0370157309002841>.

GLEICH, D. F. Pagerank beyond the web. *SIAM Review*, SIAM, v. 57, n. 3, p. 321–363, 2015.

GOLDSCHMIDT, R. R.; PASSOS, E.; BEZERRA, E. *Data mining: conceitos, ténicas, algoritmos, orientações e aplicações*. 2. ed. Rio de Janeiro: Elsevier, 2015. 276 p. ISBN 978-85-352-7822-4.

GOMEZ-URIBE, C. A.; HUNT, N. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, ACM New York, NY, USA, v. 6, n. 4, p. 1–19, 2015.

HAGBERG, A. A.; SCHULT, D. A.; SWART, P. J. Exploring network structure, dynamics, and function using networkx. In: *International Conference on Enterprise Information Systems*. Pasadena: Proceedings of the 7th Python in Science Conference, 2008. p. 11–15.

HASAN, M. A.; CHAOJI, V.; SALEM, S.; ZAKI, M. Link prediction using supervised learning. In: *SDM06: workshop on link analysis, counter-terrorism and security*. [S.l.: s.n.], 2006. v. 30, p. 798–805.

HASAN, M. A.; ZAKI, M. J. *A Survey of Link Prediction in Social Networks*. Springer US, 2011. 243–275 p. ISBN 978-1-4419-8462-3. Disponível em: <https://doi.org/10.1007/978-1-4419-8462-3_9>.

HUANG, Z.; LI, X.; CHEN, H. Link prediction approach to collaborative filtering. In: *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05)*. Denver, Colorado, USA: IEEE, 2005. p. 141–142.

JACCARD, P. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, v. 37, p. 547–579, 1901.

KASHIMA, H.; ABE, N. A parameterized probabilistic model of network evolution for supervised link prediction. In: IEEE. *Sixth International Conference on Data Mining (ICDM'06)*. [S.l.], 2006. p. 340–349.

LESKOVEC, J.; KLEINBERG, J.; FALOUTSOS, C. Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)*, ACM New York, NY, USA, v. 1, n. 1, 2007.

LIBEN-NOWELL, D.; KLEINBERG, J. The link prediction problem for social networks. In: *Proceedings of the Twelfth International Conference on Information and Knowledge Management*. New Orleans, Louisiana, USA: [s.n.], 2003.

LIBEN-NOWELL, D.; KLEINBERG, J. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, Wiley Online Library, v. 58, n. 7, p. 1019–1031, 2007.

LICHTENWALTER, R. N.; CHAWLA, N. V. Vertex collocation profiles: Subgraph counting for link analysis and prediction. In: *Proceedings of the 21st international conference on World Wide Web*. [S.l.: s.n.], 2012. p. 1019–1028.

LIN, C.-H.; KONECKI, D. M.; LIU, M.; WILSON, S. J.; NASSAR, H.; WILKINS, A. D.; GLEICH, D. F.; LICHTARGE, O. Multimodal network diffusion predicts future disease–gene–chemical associations. *Bioinformatics*, v. 35, n. 9, p. 1536–1543, 10 2018. ISSN 1367-4803.

LOFGREN, P.; BANERJEE, S.; GOEL, A. Personalized pagerank estimation and search: A bidirectional approach. In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. [S.l.: s.n.], 2016. p. 163–172.

LUTZ, M. *Programming Python*. 1. ed. [S.l.]: O'Reilly Media, Inc., 1996.

Lü, L.; ZHOU, T. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 390, n. 6, p. 1150–1170, 2011. ISSN 0378-4371. Disponível em: <https://www.sciencedirect.com/science/article/pii/S037843711000991X>.

MAGNANI, H. M. *Redes Sociais e Comunidades. Relatório Final Projeto Institucional de Iniciação Científica CNPq - IME*. Rio de Janeiro, 2015.

MARTÍNEZ, V.; BERZAL, F.; CUBERO, J.-C. A survey of link prediction in complex networks. *ACM computing surveys (CSUR)*, ACM New York, NY, USA, New York, NY, USA, v. 49, n. 4, p. 1–33, 2016.

MITZENMACHER, M. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, Taylor & Francis, v. 1, n. 2, p. 226–251, 2004.

MUNIZ, C.; GOLDSCHMIDT, R.; CHOREN, R. Combining contextual, temporal and topological information for unsupervised link prediction in social networks. *Knowledge-Based Systems*, v. 156, 2018. 21 maio de 2018.

MUTLU, E. C.; OGHAZ, T. A. Review on graph feature learning and feature extraction techniques for link prediction. *arXiv preprint arXiv:1901.03425*, 2019.

NASSAR, H.; BENSON, A. R.; GLEICH, D. F. Pairwise link prediction. In: IEEE. *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. Vancouver, Canadá, 2019. p. 386–393.

NASSAR, H.; BENSON, A. R.; GLEICH, D. F. Pairwise link prediction. *arXiv:1907.04503v1 [cs.SI]*, Julho 2019.

NASSAR, H.; BENSON, A. R.; GLEICH, D. F. Neighborhood and pagerank methods for pairwise link prediction. *Social Network Analysis and Mining*, Springer, v. 10, n. 1, p. 1–13, 2020.

NEWMAN, M. E. J. Clustering and preferential attachment in growing networks. *Physical review E*, APS, v. 64, n. 2, p. 025102, 2001.

NEWMAN, M. E. J. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, National Acad Sciences, v. 98, n. 2, p. 404–409, 2001.

NEWMAN, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, APS, v. 74, n. 3, p. 036104, 2006.

NEWMAN, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *ArXiv*, 2006.

OTTE, E.; ROUSSEAU, R. Social network analysis: A powerful strategy, also for the information sciences. *Journal of Information Science*, Sage Publications Sage CA: Thousand Oaks, CA, v. 28, n. 6, p. 441–453, 2002.

PUJARI, M. *Link Prediction in Large-Scale Complex Networks (Application to Bibliographical Networks)*. Tese (Doutorado) — Université Sorbonne Paris Cité, English, 2015.

RÜMMELE, N.; ICHISE, R.; WERTHNER, H. Exploring supervised methods for temporal link prediction in heterogeneous social networks. In: *Proceedings of the 24th international conference on world wide web*. Florence, Italy: International World Wide Web Conference Committee (IW3C2), 2015. p. 1363–1368. Disponível em: <http://dx.doi.org/10.1145/2740908.2741697.>

SONG, H. H.; CHO, T. W.; DAVE, V.; ZHANG, Y.; QIU, L. Scalable proximity estimation and link prediction in online social networks. In: *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*. Chicago, Illinois, USA: ACM, 2009. p. 322–335.

SZWARCFITER, J. L. *Teoria Computacional de Grafos*. 1. ed. Rio de Janeiro: Elsevier, 2018.

TRUDEAU, R. J. *Introduction to Graph Theory*. New York, USA: Courier Corporation, 1993.

VALVERDE-REBAZA, J.; LOPES, A. de A. Exploiting behaviors of communities of twitter users for link prediction. *Social Network Analysis and Mining*, Springer, v. 3, n. 4, p. 1063–1074, 2013.

WANG, P.; XU, B.; WU, Y.; ZHOU, X. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences*, Springer, v. 58, n. 1, p. 1–38, 2015.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. *nature*, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998.

ZHANG, J.; YU, P. S. *Link Prediction across Heterogeneous Social Networks: A Survey*. Tese (Doutorado) — University of Illinois at Chicago, Chicago, 2014.