Revista Brasileira de Computação Aplicada, Julho, 2023

DOI: 10.5335/rbca.v15i2.13817
Vol. 15, Nº 2, pp. 11–21
Homepage: seer.upf.br/index.php/rbca/index

ARTIGO ORIGINAL

# Predição de links em redes de coautoria: uma análise comparativa utilizando duas versões de métricas topológicas

# Link prediction in co-authoring networks: a comparative analysis using two versions of topological metrics

Mariana Magalhães de Mattos Coelho[1], and Claudia Marcela Justel[1]

[1]Instituto Militar de Engenharia (IME)

*mariana@ime.eb.br; cjustel@ime.eb.br

**Recebido:** 02/09/2022. **Revisado:** 12/06/2023. **Aceito:** 11/07/2023.

## Resumo

O problema denominado predição de links consiste em estimar o surgimento de arestas entre nós de um grafo que representa uma rede. Dentre as diversas abordagens do problema propostas na literatura, consideramos apenas a topológica. Utilizamos as métricas topológicas locais em duas versões: tradicional e aos pares (a última versão em duas variantes, 'ou' e 'e'). O objetivo deste trabalho é comparar quatro métricas topológicas locais, em duas versões, realizando experimentos em cinco redes reais de coautoria. Apresentamos os resultados obtidos a partir dos experimentos executados em cinco redes reais do *ArXiv*. Pelos resultados obtidos, podemos concluir que a versão aos pares obteve uma pequena vantagem em relação à versão tradicional.

**Palavras-Chave**: Análise de Redes Sociais; Aplicações de Grafos; Métricas Topológicas; Predição de Links.

## Abstract

The problem called link prediction consists of estimating the appearance of of edges between nodes of a graph representing a network. Among the different approaches of the problem proposed in the literature, we consider only the topological one. We use topological metrics in two versions: traditional and pairwise (the last version in two variants, 'or' and 'and'). The objective of this work is to compare four local topological metrics, in two versions, performing experiments in five real co-authorship networks. We present the results obtained from the experiments performed on five real *ArXiv* networks. Based on the obtained results, we can conclude that the pairwise version had a slight advantage over the traditional one.

**Keywords**: Social Network Analysis; Graph Applications; Topological Metrics; Link Prediction.

## 1 Introdução

O problema de predição de links, um problema fundamental na área de Análise de Redes Sociais, é objeto principal deste trabalho. Existem aplicações do problema em diferentes dominios, como por exemplo, predição da evolução em redes dinâmicas, indicação de novas amizades em redes sociais, recomendação de produtos e serviços. Uma rede de coautoria é representada por vértices e arestas cujos nós são os autores e as arestas são as publicações entre eles. O problema de predição de links procura identificar ligação entre pares de nós para os quais essa conexão não existe.

Para resolver esse problema, existem diferentes solu-

ções. Algumas delas utilizam características ou atributos dos nós, e outras só utilizam informação estrutural do grafo. As primeiras são conhecidas como abordagens baseadas em características, e as últimas como abordagens topológicas. Também existem abordagens que usam ambas, informação de características dos nós e informação estrutural, e são denominadas abordagens híbridas (Pujari, 2015).

O artigo de Liben-Nowell and Kleinberg (2003) introduziu diferentes métricas topológicas para resolver o problema de predição de *links*. Cada uma dessas métricas associam a um par de nós não conectados $x$, $y$ de um grafo $G$ num tempo $t$, um coeficiente, denominado $score(x, y)$. Depois disso, uma lista ordenada pelos valores de $score(x, y)$ é produzida com o objetivo de gerar um preditor de novas conexões.

Em (Nassar et al., 2019a), os autores propõem uma nova abordagem topológica para predição de *links* denominada predição aos pares que, em vez de considerar um par de nós, determina qual nó tem mais chance de formar um triângulo com uma aresta existente. Dessa forma, uma nova versão das métricas propostas por Liben-Nowell and Kleinberg (2003) é apresentada.

Em redes sociais acadêmicas, a predição de links tem sido utilizada principalmente para a predição de coautorias, atividade que indica se um par de pesquisadores poderá/irá colaborar na produção de um artigo, podendo assim otimizar a produção conjunta por meio da indicação de cientistas cujas parcerias são mais promissoras. Assim, esse tipo de predição pode ser utilizada para favorecer a comunicação entre os pesquisadores por meio da sugestão de possíveis relacionamentos, almejando potencializar o processo de produção científica. Cada vez mais as pesquisas científicas lidam com problemas complexos que, para sua resolução, exigem a colaboração de vários especialistas. A formação de equipes adequadas bem como a identificação das expertises necessárias são desafios complexos e necessários no processo da produção científica (Maruyama and Digiampietri, 2021).

O objetivo deste trabalho é fazer uma análise comparativa do desempenho de diferentes métricas topológicas, na versão tradicional e aos pares ('ou' e 'e'), em cinco redes de coautoria do *ArXiv*. Realizamos experimentos que permitiram identificar vantagens e desvantagens das duas versões tradicional e aos pares, nas variantes ('ou' e 'e'). Posteriormente, analisamos os resultados obtidos pelas duas versões. Pelos experimentos realizados, podemos concluir que a versão aos pares obteve uma pequena vantagem em relação à versão tradicional. Até o nosso conhecimento, ainda não foi realizada uma comparação das duas versões das métricas topológicas locais. Por esse motivo, analisamos a abordagem topológica das diferentes métricas nas duas versões para fins de comparação.

O restante deste artigo está organizado como se segue. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 apresenta duas abordagens topológicas. A Seção 4 apresenta a metodologia utilizada para viabilizar a comparação das métricas nas duas versões propostas pelos dois artigos mencionados anteriormente. A Seção 5 apresenta os experimentos realizados com seus respectivos resultados. A Seção 6 analisa os resultados obtidos para a comparação das métricas nos *datasets* escolhidos. Por fim, a Seção 7 apresenta a conclusão do trabalho.

## 2 Trabalhos Relacionados

Segundo Otte and 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, analisou-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.

No artigo de Liben-Nowell and Kleinberg (2003), 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. Nesse trabalho, os autores 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.

Em 2012, introduziu-se o conceito de um perfil de colocação de vértices (VCP) para fins de análise e predição de *links* topológicos (Lichtenwalter and 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 et al. (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.

Mutlu et al. (2020) 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, Nassar et al. (2019b) 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.

Assim, em 2019, Nassar et al. (2019a) propuseram prever a formação de arestas considerando um nó e uma aresta existente na rede.

Em 2021, Maruyama and Digiampietri (2021) propuseram a utilização da técnica de agrupamento e a inclusão de novos atributos que usam informações de comunidades para melhorar a previsão de relações de coautoria nas redes sociais acadêmicas.

A respeito do tipo de rede utilizada para fazer a predição de *links*, os autores a seguir analisam as características dos elementos da rede (homogênea e heterogênea). Huang et al. (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 and Kleinberg (2003), uma rede de coautoria acadêmica é analisada 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 et al. (2010) apresentaram uma abordagem diferente da proposta em (Huang et al., 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.

Em (Davis et al., 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.

Dentre os trabalhos relacionados, destacamos os dois trabalhos a seguir, ambos os quais consideram métricas topológicas em duas versões diferentes utilizando uma rede homogênea.

Os artigos de (Liben-Nowell and Kleinberg, 2003) e (Nassar et al., 2019b) consideram de formas diferentes o fechamento de triângulos ao analisar subgrafos formados por 3 nós em redes homogêneas.

## 3 Duas abordagens topológicas

Redes sociais são objetos altamente dinâmicos, que crescem e mudam rapidamente ao longo do tempo pela adição de novas ligações, de acordo com o surgimento de novas interações, na rede original. O problema de predição de *links* é um problema relacionado com a evolução da rede social ao longo do tempo. E pode ser definido da seguinte forma: dado um retrato instantâneo de uma rede social 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 and Kleinberg, 2003).

Todas as métricas utilizadas por Liben-Nowell and Kleinberg (2003) para predição de *links* associam um coeficiente a pares de nós não adjacentes $x$, $y$, denominado $score(x, y)$, a partir de um grafo de entrada e produz uma lista ordenada na ordem não crescente desses coeficientes.

Os coeficientes podem ser considerados como uma medida de proximidade ou similaridade entre um par de nós, e serão chamados neste trabalho de métricas na versão 'tradicional' . Essas métricas foram adaptadas de algumas técnicas utilizadas em Teoria de Grafos e Análise de Redes Sociais. Em geral, não foram criadas para computar a similaridade entre nós de um grafo, e, portanto, foi necessário modificá-las para o novo propósito.

A notação utilizada é a seguinte: $G = (V, E)$ um grafo não direcionado, $x \in V$ é um nó, $\Gamma(x) = \{ y \in V : (x, y) \in E \}$ é o conjunto de vizinhos do nó $x$, $| \Gamma(x) |$ é a cardinalidade do conjunto $\Gamma(x)$.

Algumas das mais populares funções de similaridade amplamente utilizadas no estado da arte (Wang et al., 2015),(Martínez et al., 2017), são as seguintes: *Vizinhos Comuns (VC)* (Newman, 2001), *Adamic-Adar (AA)* (Adamic and Adar, 2003), *Ligação Preferencial (LP)* (Newman, 2001), *Similaridade de Jaccard (JS)* (Liben-Nowell and Kleinberg, 2003).

As métricas utilizadas na versão tradicional são denotadas: Vizinhos Comuns ($VC$), Similaridade de Jaccard ($JS$), Adamic-Adar ($AA$) e Ligação Preferencial ($LP$), onde:

$$VC(x, y) = | \Gamma(x) \cap \Gamma(y) | \tag{1}$$

$$JS(x, y) = \frac{| \Gamma(x) \cap \Gamma(y) |}{| \Gamma(x) \cup \Gamma(y) |} \tag{2}$$

$$AA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log | \Gamma(z) |} \tag{3}$$

$$LP(x, y) = | \Gamma(x) | \cdot | \Gamma(y) | \tag{4}$$

Em (Nassar et al., 2019a), Nassar et al. observam que as métricas de predição de *links*, definidas em (Liben-Nowell

and Kleinberg, 2003), podem ser descritas em termos da seguinte pergunta: dado um nó $x$ na rede, quais são os nós mais propensos a serem ligados a ele? Nassar et al. consideram uma nova versão das métricas, a partir da pergunta: dada uma aresta $(u, v)$ na rede, quais são os nós mais propensos a se conectarem às extremidades da mesma, ou seja, aos vértices $u$ e $v$?

Com o objetivo de definir nova versão da proximidade ou similaridade entre cada um dos vértices que são da extremidade da aresta com o nó, que os autores denominam aos pares e denominaremos métricas na versão 'aos pares'. Será utilizada a notação a seguir: $\Gamma^*((u,v)) = \{ z \in V : \Gamma(u) \cup \Gamma(v) - \{u, v\} \}$ é o conjunto de vizinhos da aresta $(u, v) \in E$ ;

$| \Gamma^*((u,v)) |$ é a cardinalidade do conjunto $\Gamma^*((u,v))$.

As métricas adaptadas utilizando a versão aos pares definidas com os conjuntos $\Gamma^*((u,v))$ são denotadas para a versão 'aos pares' ('ou'): Vizinhos Comuns ($VC*$), Similaridade de Jaccard ($JS*$), Adamic-Adar ($AA*$) e Ligação Preferencial ($LP*$), onde:

$$VC^*(w,(u,v)) = | \Gamma(w) \cap \Gamma^*((u,v)) | \tag{5}$$

$$JS^*(w,(u,v)) = \frac{| \Gamma(w) \cap \Gamma^*((u,v)) |}{| \Gamma(w) \cup \Gamma^*((u,v)) |} \tag{6}$$

$$AA^*(w,(u,v)) = \sum_{z \in \Gamma(w) \cap \Gamma^*((u,v))} \frac{1}{log | \Gamma(z) |} \tag{7}$$

$$LP^*(w,(u,v)) = | \Gamma(w) | \cdot | \Gamma^*((u,v)) | \tag{8}$$

Num outro artigo dos mesmos autores, (Nassar et al., 2019b), as métricas adaptadas utilizando a versão aos pares definidas com os conjuntos $\Gamma^{**}((u,v))$ são denotadas para a versão 'aos pares' ('e'): Vizinhos Comuns (VC**), Similaridade de Jaccard (JS**), Adamic-Adar (AA**) e Ligação Preferencial (LP**), onde:

$\Gamma^{**}((u,v)) = \{z \in V : \Gamma(u) \cap \Gamma(v) - \{u, v\}\}$ é o conjunto de vizinhos da aresta $(u, v) \in E$ ;

$|\Gamma^{**}((u,v))|$ é a cardinalidade do conjunto $\Gamma^*((u,v))$,

$$VC^{**}(w,(u,v)) = | \Gamma(w) \cap \Gamma^{**}((u,v)) | \tag{9}$$

$$JS^{**}(w,(u,v)) = \frac{| \Gamma(w) \cap \Gamma^{**}((u,v)) |}{| \Gamma(w) \cup \Gamma^{**}((u,v)) |} \tag{10}$$

$$AA^{**}(w,(u,v)) = \sum_{z \in \Gamma(w) \cap \Gamma^{**}((u,v))} \frac{1}{log | \Gamma(z) |} \tag{11}$$

$$LP^{**}(w,(u,v)) = | \Gamma(w) | \cdot | \Gamma^{**}((u,v)) | \tag{12}$$

## 4 Metodologia

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

Seja $G = (V, E)$ um grafo no qual cada aresta $e \in E$ 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 $t$, denotada $G_t$, e que retorne uma lista de pares de elementos (dois nós não adjacentes, ou um nó e uma aresta em $G_t$, dependendo da versão da métrica usada) 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.

Para avaliar o desempenho do preditor $p$, escolheram-se os primeiros $k$ pares de predições de novas interações, ou *scores*, 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.

As redes usadas nos experimentos são redes de coautoria. O conjunto *Core* foi definido para incluir 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, *Core* = $V$, todos os autores que tiveram no mínimo uma publicação entre eles.

Os preditores utilizados foram cada uma das 4 métricas, na versão 'tradicional' e 'aos pares' ( 'e' e 'ou'). Para calcular as medidas de qualidade, foi necessário gerar a matriz de confusão, que é apresentada na Tabela 1.

**Tabela 1:** Matriz de Confusão de um Classificador - problema com 2 classes (Goldschmidt et al., 2015).

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

As medidas de qualidade da classificação que se destacam na literatura e utilizadas neste trabalho são (Zhang and Yu, 2014):

**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). A precisão está representada pela Eq. (13).

$$\text{precisão} = \frac{VP}{VP + FP} \tag{13}$$

**Acurácia**: é a fração onde o numerador é o número de verdadeiros positivos (VP) somado ao número de 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). A acurácia está representada por meio da Eq. (14).

$$\text{acurácia} = \frac{VP + VN}{VP + FP + FN + VN} \tag{14}$$

**Recall** (Revocação ou Cobertura): 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. O recall está representado por meio da Eq. (15).

$$\text{recall} = \frac{VP}{VP + FN} \tag{15}$$

**F-Mesuare** (F-1): é a média harmônica entre as medidas precisão e recall. O F-1 está representado por meio da Eq. (16).

$$\text{F-1} = \frac{2\ .(\ \text{precisão}\ .\ \text{recall})}{\text{precisão} + \text{recall}} \tag{16}$$

Além disso, também foi calculado o **preditor randômico**. Usado como base de comparação dos valores obtidos pelos outros preditores nos experimentos, faz a predição pela seleção aleatória de um par de autores que não tenham colaborado em $G_t$. A fórmula do preditor randômico é apresentada na Eq. (17):

$$\text{Pred Rand} = \frac{|E_{new}|}{|Core \times Core| - |E_{old}|}, \tag{17}$$

onde $Core = V$, $E_{new} = E_{t'} - E_t$ e $E_{old} = E_t$.

Para analisar as redes, foram determinados os seguintes valores:

**Transitividade** (transitivity): é definida pelo quociente entre o número de $K_3$ (triângulos) que existem no grafo, denotado $n_3$ e o número de pares de arestas com um vértice em comum em $G$ (possíveis $K_3$), denotado $p_3$. A Eq. (18) apresenta a fórmula para o cálculo de $T(G)$ (Newman, 2003).

$$T(G) = \frac{3\ .\ n_3}{p_3} \tag{18}$$

**Coeficiente de agrupamento médio** (*average clustering*): corresponde ao valor médio dos coeficientes de agrupamento ($c_v$) de todos os vértices $v$ do grafo $G$. A Eq. (19) apresenta a fórmula para o cálculo de $C(G)$

$$C(G) = \frac{1}{n}\sum_{v \in V} c_v, \tag{19}$$

onde $c_v = \frac{2T(v)}{|\Gamma(v)|(|\Gamma(v)|-1)}$, $T(v)$ é o número de $K_3$ no subgrafo induzido por $\Gamma(v)$ e $|\Gamma(v)|$ o grau de $v$ em $G$ (Watts and Strogatz, 1998) .

## 5 Experimentos

A liguagem de programação *Python* (LUTZ, 1996) (versão 3.9.4) foi usada para implementar o cálculo das métricas, determinar os preditores o obter os valores Top-$k$. E a biblioteca *NetworkX* (Hagberg et al., 2008) (versão 2.5.1) foi usada para a criação e manipulação de grafos.
A execução dos experimentos ocorreu em um ambiente computacional contendo o sistema operacional Ubuntu 20.04.2 LTS, com 8 núcleos de processador e 128 *Gigabytes* dememória RAM, uma máquina integrante do Laboratório de Computação de Alto Desempenho - Defesa Cibernética do IME.

Os 5 datasets considerados correspondem às redes de coautoria gr-qc, cond-mat, astro-ph, hep-ph e hep-th do *ArXiv* (`www.arxiv.org`) com informações de publicações entre os anos 1992 e 1998. Os grafos correspondentes são $G_t$ e $G_{t'}$ com $t < t'$ para os valores $t = 1997$ e $t' = 1998$. A Tabela 2 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 ($E_{new}$), respectivamente, para cada um dos cinco datasets do *ArXiv*.

**Tabela 2:** 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 1998 dos autores em $V_{1997}$ (Coelho, 2021).

| *Dataset* | $\lvert V_{1997}\rvert$ | $\lvert E_{1997}\rvert$ | $\lvert E_{1998} - E_{1997}\rvert$ | $C(G_{1997})$ | $T(G_{1997})$ |
|---|---|---|---|---|---|
| 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 |

Para comparar o desempenho das métricas na versão 'tradicional' e na versão 'aos pares' ('e' e 'ou'), os experimentos foram organizados da seguinte forma:

1. Para cada um dos 5 *datasets*, foram calculadas as 4 medidas de qualidade da classificação de cada uma das 4 métricas na versão 'tradicional' com as da versão 'aos pares' ( 'ou') e posteriormente comparados os resultados obtidos;
2. Para cada um dos 5 *datasets*, foram calculadas as 4 medidas de qualidade da classificação de cada uma das 4 métricas na versão 'tradicional' com os da versão 'aos pares' ('e') e posteriormente comparados os resultados obtidos.

Devido aos tamanhos diferentes dos *datasets* (ver Tabela 2), os valores de $k$ para estabelecer o Top-$k$ foram considerados para cada *dataset* em particular, ajustando esse valor com a quantidade de VP obtidos em cada caso.

As Tabelas 3 e 4 apresentam os resultados completos obtidos para o *dataset* gr-qc. As 8 tabelas correspondentes aos outros 4 *datasets* foram incluídas no Apêndice, e os dados consolidados da comparação são mostrados na Tabela 5 da Seção 6.

**Tabela 3:** Resultados para $G_{1997}$, $G_{1998}$ no *dataset* gr-qc do Experimento 'ou' com Pred Rand = 0,00011494

| **Top-7** | | |
|---|---|---|
| Medidas de qualidade | Métrica | |
| | *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** |

Na Tabela 3, são apresentados os resultados obtidos para o *dataset* gr-qc, onde $G_t = G_{1997}$ e $G_{t'} = G_{1998}$ para as 4 métricas. na versão 'tradicional' e 'aos pares' ( 'ou') com diferentes valores de Top-$k$. Para Vizinhos Comuns, o valor $k = 7$ foi escolhido porque, neste caso, para a versão 'tradicional' unicamente 9 valores de scores diferentes foram obtidos. Já para as métricas Similaridade de Jaccard e Adamic-Adar foi utilizado o valor $k = 25$. Para a métrica Ligação Preferencial foi utilizado o valor $k = 100$ para que, pelo menos, existisse um Verdadeiro Positivo (VP), tornando, assim, as medidas de Precisão, F-1 e Revocação diferentes de zero.

A partir dos resultados, pode-se concluir que no caso do *dataset* gr-qc, para Vizinhos Comuns e Jaccard, os resultados das medidas de qualidade da classificação obtidos pela versão 'aos pares' ('ou') foram melhores, exceto na Revocação. Para Ligação Preferencial, a versão 'aos pares' ('ou') teve melhor resultado das medidas de qualidade da classificação em todos os casos. E para Adamic-Adar, os resultados obtidos pela versão 'aos pares' ('ou') foram melhores, exceto na Acurácia. Além disso, observamos que o preditor randômico teve um valor relativamente pequeno, quando comparado com as outras medidas no *dataset* gr-qc.

**Tabela 4:** Resultados para $G_{1997}$, $G_{1998}$ no *dataset* gr-qc do Experimento 'e' com Pred Rand = 0,00011494

| **Top-7** | | |
|---|---|---|
| Medidas de qualidade | Métrica | |
| | *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 para o *dataset* gr-qc onde $G_t = G_{1997}$ e $G_{t'} = G_{1998}$ para as 4 métricas na versão 'tradicional' e 'aos pares' ( 'e') com diferentes valores de Top-$k$ são apresentados na Tabela 4. Para Vizinhos Comuns, utilizando o valor $k = 7$, os resultados das medidas de qualidade da classificação obtidos pela versão 'tradicional' foi melhor, exceto na Acurácia. Para Similaridade de Jaccard, com valor $k = 25$, o resultado foi melhor na versão aos pares para todas as medidas de qualidade, exceto na Acurácia. Para Ligação Preferencial, com valor $k = 100$, a versão 'tradicional' teve melhor resultado em todos os casos. E para Adamic-Adar, a versão 'tradicional' teve melhor resultado, exceto na Acurácia. Além disso, observamos que o preditor randômico teve um valor relativamente pequeno, quando comparado com as outras medidas no *dataset* gr-qc.

**Tabela 5:** 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** | **Acurćia** | **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** | | | | |

## 6 Análise dos Resultados

Nesta seção, são resumidos os resultados obtidos para os 5 *datasets* do *ArXiv* (gr-qc, cond-mat, astro-ph, hep-ph e hep-th) comparando o desempenho das 4 métricas ( Vizinhos Comuns, Similaridade de Jaccard, Adamic-Adar, Ligação Preferencial) nas versões 'tradicional' e 'aos pares' ('ou' e 'e'). A Tabela 5 mostra os resultados de todos os experimentos realizados indicando com * quando a métrica na versão 'aos pares' ('ou' e 'e') superou o resultado da métrica na versão 'tradicional'.

A partir dos resultados obtidos, podemos concluir que no *dataset* gr-qc, a versão 'aos pares' ('ou') teve melhor desempenho que a versão 'aos pares' ('e') , quando comparada com a versão 'tradicional' (13 contra 5). Já nos *datasets* cond-mat e astro-ph, houve empate no número de ganhos a respeito da versão 'tradicional' (7 e 6, respectivamente). E no caso do *dataset* hep-th, versão 'aos pares' na variante 'ou' também teve melhor desempenho, ainda que mais próximo, que a versão 'aos pares' na variante 'e', quando comparada com a versão 'tradicional' (5 contra 4).

A respeito das medidas de qualidade da classificação dos preditores, a Acurácia foi melhor para algumas métricas na versão 'aos pares' ('ou'), considerando todos os *datasets*. Destaca-se a métrica Vizinhos Comuns, que na versão 'aos pares' ('ou') superou o resultado da versão 'tradicional' em todos os *datasets* para a Acurácia. Já F-1 e Revocação, para a métrica Adamic-Adar na versão 'aos pares' ('ou'), superaram os resultados obtidos para a versão 'tradicional' em todos os *datasets*, com exceção do *dataset* hep-ph.

A partir dos experimentos realizados foi possível observar que a versão 'aos pares' ('e') é mais restritiva que a versão 'aos pares' ('ou'), pois para determinar seus valores, são utilizados número de elementos na interseção das vizinhanças e o número de elementos na união das vizinhanças, respectivamente. Dessa forma, podemos atribuir a esse fato que a versão 'aos pares' ('ou') teve melhores resultados em geral que a versão 'aos pares' ('e').

Além disso, constatou-se que o valor da transitividade é alto para os *datasets* gr-qc e hep-ph (ver Tabela 2). Entretanto, somente no *dataset* gr-qc a métrica na versão 'aos pares' ('ou') teve melhor resultado geral. Dessa forma, não podemos afirmar que a transitividade seja determinante para obter melhores resultados.

## 7 Conclusão

O presente trabalho se propôs a fazer uma análise comparativa das métricas topológicas locais nas versões 'tradicional' e 'aos pares' ('e' e 'ou'), utilizando a abordagem não supervisionada de predição de *links*. Foram avaliadas quatro métricas de similaridade em duas versões para 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 'aos pares' ('e)' é mais restritiva que a versão 'aos pares' ('ou'), logo, se atribui a esse fato que a versão 'aos pares' ('ou') teve melhores resultados que a versão 'aos pares' ('e'). A pequena vantagem da versão 'aos pares' ('ou') parece estar relacionada com a definição. Além disso, constatou-se, pelos experimentos, que a transitividade não parece influenciar diretamente nos resultados. Dessa forma, conseguimos obter resultados que permitem uma comparação que não existe na literatura entre as duas versões das 4 métricas topológicas.

Como possíveis trabalhos futuros, propomos realizar outros experimentos utilizando períodos de tempo diferentes para particionar 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, podem ser consideradas estruturas mais complexas para definir vizinhanças ($K_k$ para $k \geq 4$), de forma equivalente a que foi realizada por Nassar et al. (2019b) e Nassar et al. (2020) para triângulos ($K_3$) ou por Liben-Nowell and Kleinberg (2003) para arestas ($K_2$).

## 8 Apêndice

**Tabela 6:** Resultados para $G_{1997}$, $G_{1998}$ no *dataset* cond-mat do Experimento 'ou' com Pred Rand = 0,00006877

| **Top-7** | | |
|---|---|---|
| Medidas de qualidade | Métrica | |
| | *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** |

**Tabela 7:** Resultados para $G_{1997}$, $G_{1998}$ no *dataset* astro-ph do Experimento 'ou' com Pred Rand = 0,00031114

| **Top-7** | | |
|---|---|---|
| Medidas de qualidade | Métrica | |
| | *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** |

**Tabela 8:** Resultados para $G_{1997}$, $G_{1998}$ no *dataset* hep-ph do Experimento 'ou' com Pred Rand = 0,00010109

| **Top-25** | | |
|---|---|---|
| Medidas de qualidade | Métrica | |
| | *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 |

**Tabela 9:** Resultados para $G_{1997}$, $G_{1998}$ no *dataset* hep-th do Experimento 'ou' com Pred Rand = 0,0000625

| **Top-7** | | |
|---|---|---|
| Medidas de qualidade | Métrica | |
| | *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** |

**Tabela 10:** Resultados para $G_{1997}$, $G_{1998}$ no *dataset* cond-mat do Experimento 'e' com Pred Rand = 0,00006877

| **Top-7** | | |
|---|---|---|
| Medidas de qualidade | Métrica | |
| | *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 |

**Tabela 11:** Resultados para $G_{1997}$, $G_{1998}$ no *dataset* astro-ph do Experimento 'e' com Pred Rand = 0,00031114

| **Top-7** | | |
|---|---|---|
| Medidas de qualidade | Métrica | |
| | *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 |

**Tabela 12:** Resultados para $G_{1997}$, $G_{1998}$ no *dataset* hep-ph do Experimento 'e' com Pred Rand = 0,00010109

| **Top**-25 | | |
|---|---|---|
| Medidas de qualidade | Métrica | |
| | *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 |

**Tabela 13:** Resultados para $G_{1997}$, $G_{1998}$ no *dataset* hep-th do Experimento 'e' com Pred Rand = 0,0000625

| **Top**-5 | | |
|---|---|---|
| Medidas de qualidade | Métrica | |
| | *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 |

## Referências


Adamic, L. A. and Adar, E. (2003). Friends and Neighbors on the Web, *Social Networks* **25**: 211–230. https://doi.org/10.1016/S0378-8733(03)00009-1.

Benchettara, N., Kanawati, R. and Rouveirol, C. (2010). Supervised Machine Learning Applied to Link Prediction in Bipartite Social Networks, *2010 International Conference on Advances in Social Networks Analysis and Mining*, IEEE Computer Society, NW Washington, DC, United States, pp. 326–330. https://doi.org/10.1109/ASONAM.2010.87.

Coelho, M. M. d. M. (2021). *Predição de links: uma análise comparativa de métricas topológicas em redes de coautoria*, Dissertação. Mestrado em Sistemas e Computação, Instituto Militar de Engenharia, Rio de Janeiro.

Davis, D., Lichtenwalter, R. and Chawla, N. V. (2013). Supervised methods for multi-relational link prediction, *Social network analysis and mining* **3**(2): 127–141. https://doi.org/10.1007/s13278-012-0068-6.

Goldschmidt, R. R., Passos, E. and Bezerra, E. (2015). *Data mining: conceitos, ténicas, algoritmos, orientações e aplicações*, 2 edn, Elsevier, Rio de Janeiro.

Hagberg, A. A., Schult, D. A. and Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx, *International Conference on Enterprise Information Systems*, Proceedings of the 7th Python in Science Conference, Pasadena, pp. 11–15.

Huang, Z., Li, X. and Chen, H. (2005). Link prediction approach to collaborative filtering, *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05)*, IEEE, Denver, Colorado, USA, pp. 141–142. https://doi.org/10.1145/1065385.1065415.

Liben-Nowell, D. and Kleinberg, J. (2003). The Link-Prediction Problem for Social Networks, *Journal of the American Society for Information Science and Technology* **58**(7): 1019–1031. https://doi.org/10.1002/asi.20591.

Lichtenwalter, R. N. and Chawla, N. V. (2012). Vertex collocation profiles: Subgraph counting for link analysis and prediction, *Proceedings of the 21st international conference on World Wide Web*, pp. 1019–1028. https://doi.org/10.1145/2187836.2187973.

LUTZ, M. (1996). *Programming Python*, 1 edn, O'Reilly Media, Inc.

Martínez, V., Berzal, F. and Cubero, J.-C. (2017). A Survey of Link Prediction in Complex Networks, *ACM computing surveys (CSUR)* **49**(4): 1–33. https://doi.org/10.1145/3012704.

Maruyama, W. T. and Digiampietri, L. A. (2021). Combinando agrupamento e classificação para a predição de coautorias na Plataforma Lattes, *Revista Brasileira de Computação Aplicada* . https://doi.org/10.5335/rbca.v13i2.12493.

Mutlu, E. C., Oghaz, T., Rajabi, A. and Garibay, I. (2020). Review on Learning and Extracting Graph Features for Link Prediction, *Machine Learning and Knowledge Extraction* **2**(4): 672–704. https://doi.org/10.3390/make2040036.

Nassar, H., Benson, A. R. and Gleich, D. F. (2019a). Pairwise link prediction, *arXiv:1907.04503v1 [cs.SI]* . https://doi.org/10.48550/arXiv.1907.04503.

Nassar, H., Benson, A. R. and Gleich, D. F. (2019b). Pairwise link prediction, *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, Vancouver, Canadá, pp. 386–393. https://doi.org/10.1145/3341161.3342897.

Nassar, H., Benson, A. R. and Gleich, D. F. (2020). Neighborhood and Pagerank methods for pairwise link prediction, *Social Network Analysis and Mining* **10**(1): 1–13. https://doi.org/10.1007/s13278-020-00671-6.

Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks, *Physical review E* **64**(2): 025102. https://doi.org/10.1103/PhysRevE.64.025102.

Newman, M. E. J. (2003). The Structure and Function of Complex Networks, *SIAM review* **45**(2): 167–256. https://doi.org/10.1137/S003614450342480.

Otte, E. and Rousseau, R. (2002). Social Network Analysis: A Powerful Strategy, also for the Information Sciences, *Journal of Information Science* **28**(6): 441–453. https://doi.org/10.1177/016555150202800601.

Pujari, M. (2015). *Link Prediction in Large-scale Complex Networks (Application to bibliographical Networks)*, PhD thesis, Paris 13 University, France.

Rümmele, N., Ichise, R. and Werthner, H. (2015). Exploring Supervised Methods for Temporal Link Prediction in Heterogeneous Social Networks, *Proceedings of the 24th international conference on world wide web*, International World Wide Web Conference Committee (IW3C2), Florence, Italy, pp. 1363–1368. https://doi.org/10.1145/2740908.2741697.

Wang, P., Xu, B., Wu, Y. and Zhou, X. (2015). Link prediction in social networks: the state-of-the-art, *Science China Information Sciences* **58**(1): 1–38. https://doi.org/10.1007/s11432-014-5237-y.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks, *Nature* **393**(6684): 440–442. https://doi.org/10.1038/30918.

Zhang, J. and Yu, P. S. (2014). *Link Prediction across Heterogeneous Social Networks: A Survey*, PhD thesis, University of Illinois at Chicago, Chicago.