Marcelo Nunes Paolillo

Análise e Avaliação de Redes de Co-autoria para Previsão da Evolução da Rede de Colaboração do Simpósio Brasileiro IHC

Joinville

2018

UNIVERSIDADE DO ESTADO DE SANTA CATARINA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Marcelo Nunes Paolillo

ANÁLISE E AVALIAÇÃO DE REDES DE CO-AUTORIA PARA PREVISÃO DA EVOLUÇÃO DA REDE DE COLABORAÇÃO DO SIMPÓSIO BRASILEIRO IHC

Trabalho de conclusão de curso submetido à Universidade do Estado de Santa Catarina como parte dos requisitos para a obtenção do grau de Bacharel em Ciência da Computação

Rebeca Schroeder Freitas
Orientador

Joinville, Junho de 2018

ANÁLISE E AVALIAÇÃO DE REDES DE CO-AUTORIA PARA PREVISÃO DA EVOLUÇÃO DA REDE DE COLABORAÇÃO DO SIMPÓSIO BRASILEIRO IHC

Marcelo Nunes Paolillo

Este Trabalho de Conclusão de Curso foi julgado adequado para a obtenção do título de Bacharel em Ciência da Computação e aprovado em sua forma final pelo Curso de Ciência da Computação Integral do CCT/UDESC.

Banca Examinadora

Rebeca Schroeder Freitas - Doutora (orientador)

Avanilde Kemczinski - Doutora

Isabela Gasparini - Doutora

Resumo

Redes sociais são estruturas compostas por pessoas, conectadas por um ou vários tipos de relações, que podem ser representadas por estruturas conhecidas como grafos. A previsão de relacionamentos em uma rede social é uma tarefa complexa, dado o grande número de aspectos estruturais e externos com os quais está vinculada. O estudo desses aspectos pode ser usado em redes sociais acadêmicas para melhorar ou maximizar as colaborações, indicando potenciais parcerias para o desenvolvimento de um projeto ou co-autores para publicar trabalhos. Este projeto visa abordar o reconhecimento de padrões inerentes à evolução de redes sociais, mais especificamente a Rede de Colaboração do Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (IHC) ao longo do período de 1998 a 2018. Abordagens para previsão de novas co-autorias serão aplicadas, combinando questões estruturais e topológicas de grafos com técnicas de aprendizado de máquina. Assim, espera-se encontrar características ou padrões importantes para projetar um modelo computacional que consiga prever novas co-autorias na rede de colaboração do IHC. Como resultado geral, pretende-se contribuir para a área de predição de arestas em redes complexas.

Palavras-chaves: grafos, *machine learning*, redes sociais, redes acadêmicas, *link prediction*.

Abstract

Social networks are structures representing people connected to each other by one or several types of relationships, which can be represented as graphs. The prediction of relationships in a social network is a complex task given the large number of structural and external aspects involved. The study of these aspects can be useful for academic social networks to improve or maximize collaborations, indicating potential partnerships for the development of a project or co-authors to publish papers. This project aims to address the recognition of patterns in the evolution of social networks, more specifically the Collaboration network of the *Brazilian Symposium on Human Factors in Computational Systems* (IHC) from 1998 to 2018. In order to predict new co-authors, this work will combine approches using structural and topological features on graphs with machine learning techniques. We expect to find important characteristics or patterns to design a computational model to predict new co-authorships for the collaboration network of IHC. As a general result, we hope to contribute to the link prediction area for complex networks.

Keywords: graphs, machine learning, social networks, academic networks, link prediction.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AA	*Índice de Adamic/Adar*
ARS	*Análise de Redes Sociais*
AUC	*Area Under Curve*
CM	*Caminho Mínimo*
CN	*Common Neighbours*
CS	*Cosine Similarity*
DT	*Árvore de Decisão*
FL	*FriendLink*
HDI	*Hub Depressed Index*
HPI	*Hub Promoted Index*
JC	*Coeficiente de Jaccard*
KNN	*K-Nearest Neighbors*
LHN	*Leicht Holme Newman Index*
LPI	*Local Path Index*
LRW	*Local Random Walk*
NB	*Naive Bayes*
PA	*Preferential Attachment*
RA	*Resource Allocation*
RF	*Random Forest*
RN	*Rede Neural*

RW	*Random Walk*
SR	*SimRank*
SVM	*Support Vector Machine*

1 Introdução

Com o advento e popularização da Web para uso pessoal na década de noventa, a quantidade de dados disponíveis aumentou de maneira considerável. Esses dados se relacionam de diversas maneiras, e a forma comum de representação destes relacionamentos é por meio da criação de redes, sejam elas redes de computadores, redes organizacionais, redes de relações empresariais, redes neurais, redes metabólicas, redes de distribuição, redes de vasos sanguíneos, rotas de entrega postal, redes de relacionamento interpessoal ou sociais, e inúmeras outras. (NEWMAN, 2003; LESKOVEC; KLEINBERG; FALOUTSOS, 2006).

No âmbito da matemática e da computação, redes podem ser definidas formalmente por estruturas conhecidas como grafos, $G = (V, E)$, onde: V corresponde ao conjunto de vértices ou nós e E compreende o conjunto de arestas que conectam os vértices. Um exemplo de aresta e pode ser representado por $e = (v_1, v_2)$, tal que $e \in E$, $v_1 \in V$ e $v_2 \in V$. Grafos muitas vezes são usados para modelar sistemas reais, nessas instâncias a estrutura não se comporta de maneira regular, mas também não de maneira aleatória, apresentando padrões topológicos sutis ao longo de sua evolução do tempo. Tais estruturas quando usadas neste sentido são caracterizados como redes complexas. As publicações de Watts e Strogatz (1998) e Barabási et al. (2002) expandiram a noção dessas redes para duas subclasses, redes de pequeno mundo e redes livres de escala. Estas subclasses são resultados da identificação das primeiras propriedades padrão presentes em grafos que modelam sistemas reais, portanto foram o ponta-pé inicial para o início dos estudos na área.

A partir disso, os estudos subsequentes buscaram definir novos conceitos e medidas para caracterizar a topologia das redes reais. Estes estudos foram motivados pela expectativa de que a compreensão e modelagem da estrutura de uma rede complexa levaria a um melhor conhecimento de seus mecanismos evolutivos, e a um melhor entendimento de seu comportamento dinâmico e funcional (BOCCALETTI et al., 2006). Redes que modelam sistemas reais são dinâmicas, e mudam e crescem rapidamente ao longo do tempo através da criação de novas arestas e nodos. Entender os mecanismos pelos quais elas evoluem define o problema da predição de arestas (*the link prediction problem*). Isto é, dado o estado de uma rede em um momento t, busca-se prever com precisão arestas

que serão adicionadas a rede durante um intervalo de tempo de t até um momento futuro t' (LIBEN-NOWELL; KLEINBERG, 2004).

A aplicação de técnicas de predição de arestas é vista normalmente em sistemas de recomendação, como sugestões para produtos relacionados, ou em redes sociais para sugestões de amizade. Por outro lado, estas podem ser aplicadas em diversos outros tipos de redes, não só para prever novas conexões, mas também para outros casos de uso: estudo de redes terroristas (HASAN; SALEM; ZAKI, 2006); conexões em microblogs como *Twitter* (YIN; HONG; DAVISON, 2011), (YANTAO et al., 2013); previsão de padrões de uso de *websites* com redes de *hyperlinks* (ZHU; HONG; HUGHES, 2002); recomendação de reagentes químicos (SAVAGE et al., 2017); controle de privacidade em redes sociais (AL-OUFI; KIM; El Saddik, 2011); sumarização de documentos (FENG et al., 2012); para aumentar o foco de campanhas de marketing (GRYLLOS; MAKRIS; VIKATOS, 2017); e inúmeros outros.

Este trabalho busca abordar o problema de predição de arestas em redes colaborativas. Estas são redes que representam grupos ou conjuntos de pessoas que buscam trabalhar de maneira colaborativa ou cooperativa, com o intuito de coordenar esforços a fim de atingir metas em comum (NEWMAN, 2003). Um tipo de rede colaborativa é a rede de colaboração científica, que é formada com um viés acadêmico, por professores, pesquisadores e estudantes. Neste contexto, o objetivo é a busca por novos conhecimentos, ou investigar os conhecimentos existentes, de maneira a contribuir com alguma área de interesse.

A recomendação de colaborações que ainda não ocorreram é importante para o desenvolvimento do grupo de pesquisa. Novas colaborações apresentam benefícios, uma vez que trazem experiência ou novos conhecimentos que podem contribuir com o estudo de uma devida área, o que leva a novas publicações. Publicações influentes podem melhorar a reputação e reconhecimento dos autores, o que aumenta a possibilidade de patrocínio que propicia novas e maiores pesquisas (LI; LIAO; YEN, 2013). Além disto, abre a possibilidade de novas parcerias, devido a similaridades anteriormente desconhecidas, que levam a novas colaborações.

A abordagem do problema pode ser vista para o caso de redes colaborativas, em geral, de duas formas: abordagens por medida de similaridade e abordagens com aprendizado de máquina (GAO et al., 2014). A primeira tende a usar medidas de si-

milaridade entre os pares de indivíduos para ranquear os pares e realizar previsões com base em notas de corte no ranqueamento. Já com aprendizado de máquina, em geral, se transforma o problema em classificação binária. Neste caso, são escolhidos um ou mais classificadores ou modelos probabilísticos, que são treinados com múltiplas medidas de similaridade e atributos dos vértices (HASAN; SALEM; ZAKI, 2006), para prever se um par de vértices pertence a classe negativa, se os autores não colaboram, ou positiva, se os autores colaboram.

Apesar dos métodos baseados em classificadores parecerem mais atrativos, devido ao uso de múltiplas características, estes métodos também trazem as suas dificuldades. Como por exemplo, o extremo desbalanceamento de classe, onde ocorre que o número total de uma classe é muito maior que outra. Isso ocorre em no problema em questão pois, em geral, o número de arestas possíveis é quadrático ao número de vértices no grafo, enquanto o número de arestas reais é apenas uma pequena fração deste número. Por consequência, classificadores têm dificuldade de prever a existência ou não de uma aresta no futuro, devido a sub-representação da classe de interesse. Classificadores acabam por reconhecer que na grande maioria dos casos não há arestas entre dois vértices, logo tendem a prever que não há arestas sempre (HASAN; ZAKI, 2011).

O foco deste projeto é avaliar a rede de co-autoria do Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (IHC) ao longo do período de publicações de 1998 a 2018. O objetivo é sugerir ou recomendar novas colaborações, usando uma combinação de características estruturais da rede e características específicas dos autores e publicações. Para apoiar este processo serão aplicados algoritmos de aprendizado de máquina, considerando as características que obtêm maior taxa de acerto. Ou seja, características onde, dada uma previsão de co-autorias sobre um grafo em um intervalo de tempo, no qual sabe-se o resultado previamente, há maior relação entre o número de arestas previstas corretamente e o número de arestas reais. Dessa forma, auxiliam no desenvolvimento da comunidade, visto que encontram fatores que levam a colaboração entre autores, logo contribuem com a área de predição de arestas no geral.

1.1 Objetivos

Esta seção apresenta o objetivo geral do trabalho e os objetivos específicos, por meio dos quais o objetivo geral seja concretizado.

1.1.1 Objetivos Gerais

Este trabalho visa colaborar com o estudo e avaliação de abordagens e métodos para a previsão da evolução de redes sociais. Com o enfoque em redes de co-autoria, o objetivo geral deste trabalho é identificar o modelo de evolução da rede do Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (IHC), durante os últimos 20 anos de suas publicações. O modelo de evolução é caracterizado por premissas que determinam novas conexões ao longo dos anos, utilizando para isto técnicas de predição de arestas em grafos.

1.1.2 Objetivos Específicos

- Analisar e avaliar à evolução da rede, identificando características importantes do grafo que possam estar associadas a evolução da mesma;
- Aplicar métodos encontrados para previsão da evolução em intervalos de anos específicos do IHC, gerando novas redes de colaboração;
- Comparar as redes de colaboração geradas pelos métodos de previsão com as redes de co-autoria do IHC dos anos subsequentes;
- Computar a assertividade dos métodos de previsão e identificar seus respectivos modelos de evolução. Espera-se com isto selecionar e recomendar os modelos e respectivos métodos que obtiveram a maior assertividade.

1.2 Metodologia

O trabalho inicia-se com a etapa de revisão sobre os principais conceitos envolvidos no tema, como: questões estruturais e métricas para grafos; o estudo de fatores que influenciam na evolução de redes sociais; e ainda técnicas e abordagens utilizadas na previsão de

conexões.

Simultaneamente, deve-se iniciar a atualização da base de dados MySQL até o ano de 2017, visto que no momento tem-se apenas dados de 1998 a 2015. Estes dados já estão disponíveis, portanto a tarefa se dará pela aplicação de scripts em Python para a adequação destes novos dados à estrutura na qual o banco se encontra.

Esta atualização permite gerar a rede de colaboração a partir da base do MySQL e representá-la através de uma estrutura em grafo, o que ocorrerá em sequência.

Após a revisão conceitual e de trabalhos relacionados, e atualização das bases de dados, serão discutidas as características que serão representadas no grafo da rede de co-autoria. Estas características são fundamentais, pois definem que tipo de previsão será possível através do uso do modelo, bem como implicam nas características a serem aplicadas no mesmo.

Seguido disso, serão avaliados ambientes para a visualização de grafos. Com o intuito de tornar visual a evolução da rede de co-autoria ao longo dos anos, e ao mesmo tempo permitindo a percepção de padrões ligados a esta.

Subsequentemente também serão escolhidos os métodos, métricas, e características para a implementação do algoritmo para a previsão de novas conexões. Dos quais, diversos serão aplicados para avaliação e comparação de eficiência e eficácia na tarefa citada.

1.3 Estrutura do Trabalho

De acordo com o objetivo deste estudo, o trabalho foi estruturado em cinco capítulos. O Capítulo 2 descreve os fundamentos envolvidos neste trabalho, apresentando mais formalmente os conceitos de alguns termos usados. Iniciando-se com o conceito de redes complexas e suas subclasses, seguido de redes de colaboração científica, para que então sejam apresentadas teorias sociais usadas como bases para geração de medidas de predição de arestas. Abordagens, métricas e ferramentas para resolução de problemas de predicção de arestas são apresentados no Capítulo 3. O Capítulo 4 apresenta os trabalhos que estão relacionados ao tema desta pesquisa, e que auxiliaram de alguma forma para utilizar como base para esse estudo. No Capítulo 5 é descrita a proposta deste trabalho, onde também é mostrado todas as etapas que foram realizadas desde o início deste trabalho. Por fim, no

Capítulo 6 são apresentadas as conclusões parciais e o cronograma dos trabalhos futuros.

2 Fundamentação Teórica

Este capítulo está dedicado aos conceitos e termos necessários para o entendimento deste trabalho. O conceito de grafos complexos é inicialmente apresentado, o qual inclui as redes de colaboração e co-autorias científicas. Em seguida, serão analisadas as teorias relacionadas à evolução de redes, para então apresentar o problema de previsão de arestas e sua importância para o estudo de redes de colaboração. Por fim, são descritas algumas ferramentas utilizadas na abordagem do problema e algumas considerações.

2.1 Redes Complexas

Grafos são estruturas matemáticas usadas para representar relacionamentos, e que podem ser definidas por: $G = (V, E)$, onde: V corresponde ao conjunto de vértices, também conhecidos como nós ou nodos. Nós podem ser conectados por arestas ou ligações, como por exemplo $e = (v_1, v_2)$, onde $v_1 \in V$, $v_2 \in V$ e, $e \in E$ representando um relacionamento. Nessa instância, o grafo é considerado não direcionado, ou seja arestas deste não apontam para nenhuma direção. Tais estruturas podem ser usadas para modelar vários tipos de relações e processos do mundo real, seja para questões físicas, químicas, biológicas, ou sociais. Quando são aplicadas nestes contextos, são associadas ao termo *rede*.

Redes complexas são estruturas que apresentam características diferentes de grafos gerados aleatoriamente, onde lê-se aleatórios, ou grafos regulares, ou grafos gerados de acordo com algum padrão fixo. Essas características refletem na topologia da rede, com padrões de conexão entre elementos não retratando completamente grafos aleatórios, nem grafos regulares, mas sim algo intermediário (KIM; WILHELM, 2008).

O estudo de redes complexas busca compreender a interação de padrões topológicos presentes na evolução dessas estruturas, para auxiliar na avaliação ou tomada de decisão associados a sistemas reais cujos relacionamentos podem ser modelados de acordo com grafos. O estudo destes padrões é de importância para várias questões práticas, como por exemplo na sociologia, para melhor entendimento sobre os mecanismos envolvidos na formação de comportamentos sociais coletivos, como novos hábitos, moda, e mudança de

opinião. Já na medicina, existe evidência da existência de que algumas doenças cerebrais são resultados de comportamento anormal em neurônios, que podem ser identificados de maneira mais eficiente, quando se sabem quais padrões são comuns e quais não são (BOCCALETTI et al., 2006).

Algumas propriedades de redes complexas foram identificadas, e deram origem as subclasses *redes de pequeno mundo* e *redes livres de escala*. Estas subclasses não são mutualmente exclusivas, inclusive várias redes complexas apresentam ambas as propriedades em si. As próximas seções apresentam as referidas subclasses.

2.1.1 Redes de pequeno mundo

As redes de pequeno mundo (*Small World Networks*) são redes cuja maioria dos nodos não são vizinhos um do outro, porém os vizinhos de qualquer nodo são provavelmente vizinhos entre si. Portanto, a maioria dos vértices pode ser alcançado a partir de qualquer vértice na rede através de um número relativamente pequeno de passos. Esta classe de redes foi inicialmente identificada por Watts e Strogatz (1998), e pode ser mais formalmente definida como uma rede onde espera-se que a distância média L entre dois vértices aleatoriamente selecionados, cresce proporcionalmente ao logarítmo do número de vértices N na rede.

O mesmo estudo também mostrou que grafos podem ser classificados de acordo com duas características estruturais independentes. Sendo elas o coeficiente de agrupamento (C), que é a medida do quanto os nodos de um grafo tendem a se aglomerar ou se aproximar um do outro, e a distância média entre vértices (L), também conhecido como a média de caminhos mínimos. A ideia por trás de uma rede de pequeno mundo é que a primeira dessas características tenha um valor alto e a segunda um valor pequeno, isto é, proporcional ao logaritmo do número de nodos na rede.

Duas versões desta medida de agrupamento existem: o global e o local. A versão global foi projetada para fornecer uma indicação geral do agrupamento na rede, enquanto o local fornece uma indicação do quanto a vinhança de um nós está próxima a formar um clique, ou grafo completo, em que todas as arestas entre os vértices estão presentes.

Watts e Strogatz (1998) definem o coeficiente de agrupamento local c_i^{ws} de um

nodo i da seguinte maneira:

$$c_i^{ws} = \frac{2E_i}{k_i(k_i - 1)}$$

Onde E_i é o número de arestas entre os vizinhos de i, e k_i o número de arestas incidentes ao vértice i, ou grau. Vale relembrar que as medidas se referem a grafos não direcionados.

A medida global, também conhecida como transitividade, definida por Luce e Perry (1949):

$$C^{\Delta} = \frac{n^o triplas fechadas}{n^o \, total \, de \, triplas}$$

Onde, uma tripla é formada por três nós conectados por dois laços (tripla aberta) ou três (tripla fechada). Um triângulo possui três triplas fechadas, uma centrada em cada um dos nós.

Uma alternativa para obter o coeficiente de agrupamento global foi proposta por Watts e Strogatz (1998), que é a média da medida local para todos os vértices, ou seja:

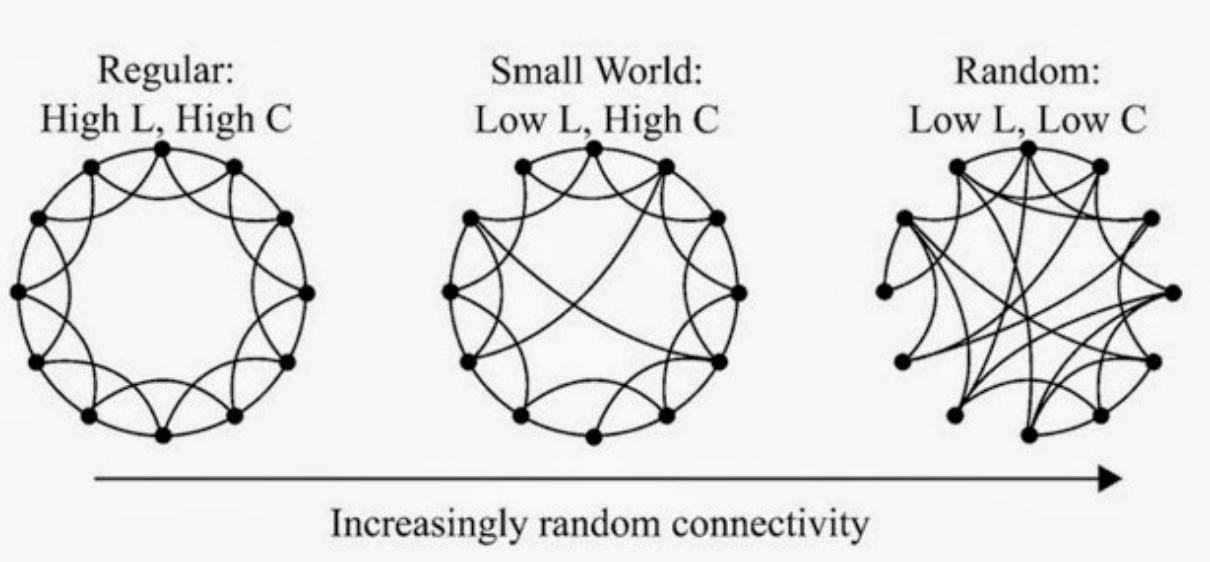
$$C_i^{ws} = \frac{1}{n} \sum_{i=1}^{n} c_i^{ws}$$

Pode-se quantificar o quanto uma rede é de pequeno mundo através do coeficiente *small* (σ), definido pela comparação do agrupamento, da distância média entre dois vértices na rede em questão, e uma rede equivalente aleatória com o mesmo grau médio (HUMPHRIES; GURNEY, 2008):

$$\sigma = \frac{\frac{C}{C_r}}{\frac{L}{L_r}}$$

Onde, C e C_r são os coeficientes de agrupamento da rede em questão e da rede aleatória, respectivamente. Da mesma forma, L e L_r são a distância média entre dois vértices escolhidos aleatoriamente na rede, para a rede em questão e para uma rede gerada aleatoriamente. Nota-se que se $\sigma > 1$, a rede é de pequeno mundo.

Figura 2.1: Diferença entre um grafo regular, rede de pequeno mundo e grafo aleatório



Fonte: (WATTS; STROGATZ, 1998)

Redes de pequeno mundo tendem a ter baixo tamanho de caminho médio, L, e alto coeficiente de agrupamento C, o que as diferencia de grafos regulares que apresentam alto L e alto C. Em contrapartida, grafos aleatórios possuem baixo L e baixo C, conforme demonstrado pela Figura 2.1 (WATTS; STROGATZ, 1998).

2.1.2 Redes Sem Escala

Redes sem escala, ou redes livres de escala (*Scale-Free Networks*), são redes em que a distribuição de grau, que corresponde a proporção de vértices de determinado grau no grafo, tende a seguir a lei de potência. Tal lei pode ser observada se em dado relacionamento, uma mudança relativa em uma quantidade resulta em uma alteração relativa proporcional na outra quantidade. Ou seja, neste caso, a maioria dos vértices tem poucas conexões, mas uma quantia proporcionalmente menor de vértices mostra um número mais elevado de ligações. Isso tende a resultar em conexões mais frequentes entres nodos de maior grau, e de maneira geral, a probabilidade de um nodo se ligar a outro nodo é diretamente proporcional a seu grau.

Um dos primeiros estudos que identificou esta propriedade foi Barabási e Albert (1999), que ao mapear a topologia de uma porção da *World Wide Web*, observou que um número pequeno de nodos possuiam bem mais conexões que outros, e dessa forma a distribuição de grau respeitava a lei de potência.

Li et al. (2005), propuseram uma métrica formal para a definição do nível de escala de uma rede:

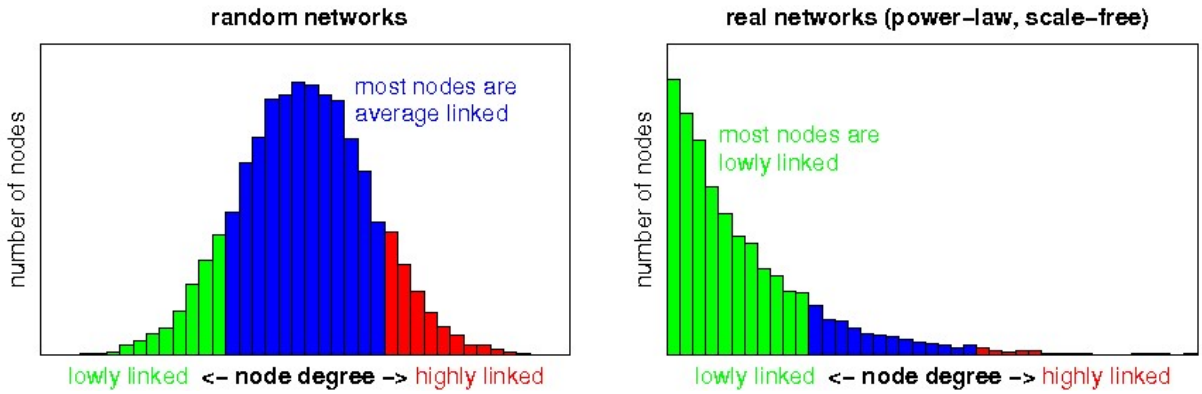
$$s(G) = \sum_{(u,v)\in E} k(u).k(v)$$

Sendo u e v vértices, e $k(u)$ e $k(v)$ o grau dos mesmos, e:

$$S(G) = \frac{s(G)}{s_{max}}$$

Com s_{max} sendo o maior valor de $s(H)$ e H o conjunto de todos os grafos com a distribuição de grau idêntica a G. Esta métrica resulta em valores entre 0 a 1, onde um grafo com S(G) = 1 é um grafo livre de escala. Isso significa que uma pequena quantidade de vértices domina a rede, com um número muito maior de conexões do que a grande maioria dos outros vértices, este fenômeno é comummente associado a desigualdade econômica descrita pela expressão: "*The rich get richer, and the poor get poorer*"(Os ricos ficam mais ricos e os pobres ficam mais pobres) (BARABÁSI; ALBERT, 1999).

Figura 2.2: Distribuição de grau de uma grafo aleatório e para uma rede livre de escala



Fonte: (SCHOLZ, 2018)

A Figura 2.2 mostra a diferença na distribuição de grau entre um grafo gerado aleatoriamente e um grafo que modela uma sistema real, que possui a propriedade de ser livre de escala. As barras verdes representam a contagem de vértices com baixo grau, as barras em azul representam vértices com grau intermediário, e as vermelhas representam vértices com grau mais alto.

Com base nisso, Barabási e Albert (1999) propuseram associar *Preferential*

Attachment, ou conexão preferencial, à evolução de redes. Isto é, a probabilidade de uma nova conexão a um vértice na rede, é diretamente proporcional ao número de vértices com os quais este já está conectado.

2.2 Redes de Colaboração Científica

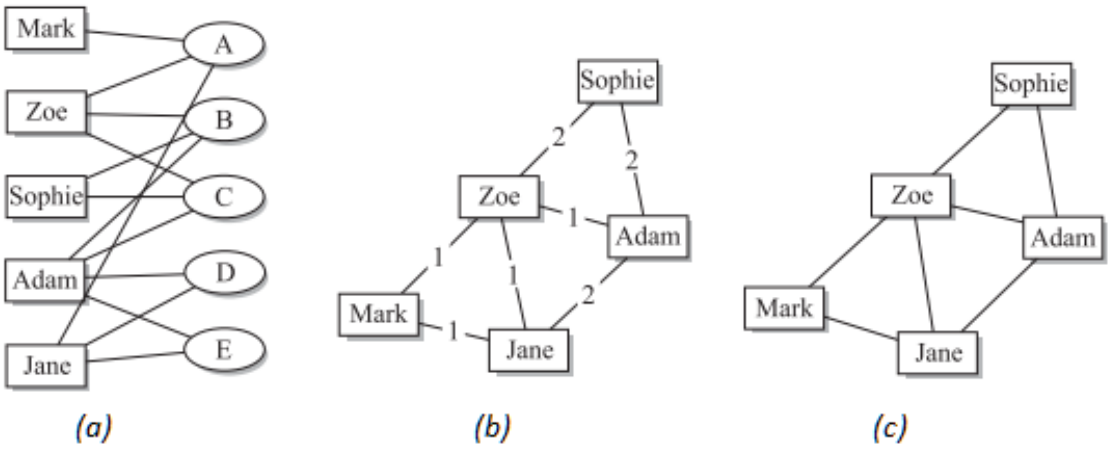
Redes complexas podem ser vistas em uma variedade de sistemas reais, como redes sociais. Os nodos de uma rede social representam pessoas ou outras entidades em um contexto social. Já as arestas representam interação, amizade, colaboração, ou influência entre estas entidades.

Redes colaborativas, por sua vez, são redes que representam grupos ou conjuntos de pessoas que buscam trabalhar de maneira colaborativa ou cooperativa, com o intuito de coordenar esforços a fim de atingir metas em comum (NEWMAN, 2003). Um tipo de rede colaborativa é a rede de colaboração científica, que é formada com um viés acadêmico, por professores, pesquisadores e estudantes. Neste contexto, o objetivo é a busca por novos conhecimentos, ou investigar os conhecimentos existentes, de maneira a contribuir com alguma área de interesse.

Dentre as variantes de redes de colaboração científica, estão as de colaboração indireta, conhecidas como redes de citação, e as de colaboração direta, na forma de redes de co-autoria. O foco deste trabalho está em redes de co-autoria, que são redes cujos vértices são autores de artigos científicos, e as arestas ou relacionamentos entre tais vértices representam publicações em conjunto (NEWMAN, 2001).

Esse tipo de rede pode ser estruturalmente vista de diversas formas. A Figura 2.3 ilustra os três tipos de representação de redes de co-autoria mais comuns. Pode-se observar a esquerda um grafo de colaboração bipartido (a), onde Mark, Zoe e Jane publicaram o artigo A. Os grafos seguintes correspondem a mesma rede em um grafo ponderado (b), onde o número de publicações conjuntas é o peso nas arestas, e um grafo não ponderado (c) onde o número de publicações conjuntas não é visível (GUNS, 2016).

Figura 2.3: Rede de colaboração (a) bipartida, (b) ponderada e (c) não-ponderada



Fonte: (GUNS, 2016)

A colaboração entre as entidades pertencentes a uma rede de colaboração científica, ocorre através de uma orientação discussão de ideias, ou até participação em um projeto de pesquisa. Porém, não é apenas isso que acarreta em colaborações. Mendonça (2017), realizou um breve levantamento de fatores que levam a colaboração, e estes são descritos a seguir:

- **Colaboração de formação**: que é a colaboração entre um orientador e um orientando, ela acontece devido a necessidade de uma contribuição especializada de um orientador para alcançar os objetivos da pesquisa, que o orientado necessita para adquirir conhecimento e habilidade.

- **Colaboração teórica e experimental**: que são tipos de colaboração de acordo com o objetivo da publicação. Publicações teóricas tendem a ter menos colaboração, do que as publicação de cunho experimental (BALANCIERI, 2004).

- **Proximidade de colaboração**: Está associada a maior facilidade de colaboração entre pesquisadores próximos. A internet contribuiu com este fator, mas a proximidade também pode estar ligada a questões culturais, idioma, interesses e oportunidades (KATZ; MARTIN, 1997).

- **Produtividade colaborativa**: autores produtivos no âmbito de número de publicações, tendem a colaborar mais (KATZ; MARTIN, 1997).

- **Tamanho do grupo de pesquisa**: áreas de pesquisa com maior número de autores tendem a atrair mais colaborações novas, do que áreas de pesquisa pouco exploradas. O que leva a criação de grupos com objetivos em comum (NARIN; STEVENS; WHITLOW, 1991).

- **Interdisciplinaridade**: acontece da fusão de conhecimentos de áreas consideradas anteriormente distintas, que propiciam avanços significativos quando em conjunto.

- **Compartilhamento de Recursos**: A necessidade de compartilhar equipamentos caros e complexos motiva a colaboração entre pesquisadores (KATZ; MARTIN, 1997).

- **Busca por reconhecimento**: autores relativamente desconhecidos buscam trabalhar com aqueles que são conhecidos para alcançar reconhecimento (NARIN; STEVENS; WHITLOW, 1991).

A recomendação de colaborações que ainda não ocorreram é importante para o desenvolvimento do grupo de pesquisa. Novas colaborações apresentam benefícios, uma vez que trazem experiência ou novos conhecimentos que podem contribuir com o estudo de uma devida área, o que leva a novas publicações. Publicações influentes podem melhorar a reputação e reconhecimento dos autores, o que e aumenta a possibilidade de patrocínio que propicia novas e maiores pesquisas (LI; LIAO; YEN, 2013). Além disto, abre a possibilidade de novas parcerias, devido a similaridades anteriormente desconhecidas, que levam a novas colaborações.

A evolução de redes de co-autoria ao longo do tempo refletem em padrões de grupos ou áreas de pesquisa. Assim, um experimento possível seria separar a rede em intervalos de tempo e observar padrões que permitem descobrir fatores correlacionados à formação de novas colaborações. A partir do momento em que se sabe estes fatores, a recomendação de novas colaborações se torna viável. Uma abordagem para tal experimento é a utilizada no problema de previsão de arestas, que consiste em prever arestas que se formaram em um intervalo de tempo, e cruzar com os resultados da realidade, através do uso de medidas baseadas em teorias sociais.

2.3 Teorias Sociais

Definida a importância dos aspectos que levam a colaboração, busca-se apresentar as teorias que levaram a criação de medidas para a previsão de arestas, Li, Fang e Sheng (2015) constatam que entender as bases teóricas por trás do estudo de recomendação ou previsão de arestas possui três benefícios claros: (i) ajudam a entender por quê e em quais circunstâncias o método utilizado funciona; (ii) ajudam a identificar limitações de métodos existentes, e por consequência, desenvolver métodos melhores; (iii) teorias nos informam sobre fatores genéricos que afetam conexões, e que podem ser usados para projetar novos e mais eficientes métodos de recomendação.

Observações que inspiraram técnicas de recomendação podem ser também separadas, de modo geral, em três grupos. Mais especificamente, é observada a formação de conexão entre duas entidades sociais de acordo com: grau de similaridade entre si; conexões e decisões de seus vizinhos; e caminhos que os conectam na rede. Essas observações são respondidas, respectivamente, pela teoria da homofilia (MCPHERSON; SMITH-LOVIN; COOK, 2001), teoria da interação social (BECKER, 1974), e teoria do balanço cognitivo (HEIDER, 2013). Uma breve descrição dessas se dá nas seguintes subseções, nota-se que as estratégias para resolução do problema de predição de arestas descritas no próximo capítulo, usam das medidas inspiradas nestas teorias para resolução do problema.

2.3.1 Teoria da Homofilia

A teoria proposta por McPherson, Smith-Lovin e Cook (2001 apud LI; FANG; SHENG, 2015), constata que o contato entre pessoas semelhantes ocorre mais do que entre pessoas não-semelhantes. E, teoricamente, o fluxo de informação de uma pessoa a outra é uma função decrescente a distância entre as duas num contexto socio-demográfico. Portanto, pessoas, ou vértices num grafo, têm menos chances de se conectarem com aqueles que, de acordo com dada medida de similaridade, são pouco parecidos. A teoria ainda é expandida pela noção de que pessoas criam seu "mundo social"de acordo com as pessoas com que elas decidem interagir, isso faz com que sejam criados nichos sociais distintos. E, por isso, implica que a similaridade contribui com o estabelecimento de novas conexões.

A ideia de que similaridade gera conexões inspirou diversas medidas, como as de similaridade: de perfil, semântica, geográfica, demográfica, interesse social e outras

mais. Estas medidas podem ser usadas em modelos com intuito de apontar que quanto mais semelhantes autores são, maior a chance de colaboração no futuro.

2.3.2 Teoria da Interação Social

De acordo com a teoria de interação social proposta por Becker (1974 apud LI; FANG; SHENG, 2015), quando uma entidade social toma uma decisão, a mesma depende das decisões de seus vizinhos sociais. Mais especificamente, uma decisão social, como uma conexão, é uma decisão que influencia e é influenciada pelas decisões daqueles que são próximos socialmente do decisor. Num contexto mais geral, uma entidade depende de informação, neste caso, informações sobre o que outros estão pensando em fazer ou já fizeram para realizar a sua própria decisão.

A noção desta teoria é expandida em particular para duas questões. Na primeira, uma entidade social tende a se estabelecer em grupo por suas decisões estarem de acordo com as entidades sociais presentes no grupo. Já na segunda, os membros de um grupo tendem a dissipar informações aos demais membros, dessa forma influenciando-os em suas decisões. Por isso, é utilizada como base para alguns métodos de recomendação ou predição de arestas, como por exemplo, medidas de semelhança baseadas em vizinhança.

2.3.3 Teoria do Balanço Cognitivo

A teoria do balanço cognitivo proposta por Heider (2013 apud LI; FANG; SHENG, 2015), é explicada com uma perspectiva psicológica do fenômeno da transitividade de redes, que diferencia redes sociais de outros tipos de redes. Este fenômeno corresponde a questão social de que entidades sociais que estão indiretamente associadas, podem vir a se associar no futuro. De acordo com a teoria, sentimentos ou atitudes de entidades sociais indiretamente associadas podem gradualmente se tornar consistentes, o que pode levar estas a se conectar no futuro.

Este teoria é usada como base para métodos de recomendação ou previsão baseados nos caminhos entre dois vértices como Katz, bem como métodos probabilísticos de caminhada aleatória. As quais usam de caminhos indiretos e estocasticidade para modelar relacionamentos que podem vir a surgir de maneira indireta.

2.4 Considerações sobre o Capítulo

Neste capítulo foram apresentados os conceitos básicos relacionados a redes complexas, isto é, redes que modelam sistemas reais. Tais redes costumam apresentar características de redes de pequeno mundo e/ou redes sem escala, que por sua vez estão presentes em redes colaborativas de co-autoria, o foco deste trabalho. Conforme discutido, ambas são de grande importância para o entendimento da evolução de uma comunidade científica, bem como para a recomendação de novas colaborações. Por esta razão, foram expostos alguns dos fatores que levam a colaboração em um contexto mais geral, seguida da fundamentação das teorias sociais que ligam estes fatores a medidas topológicas do grafo gerado pelas redes. No próximo capítulo é formalizado o problema de predição de arestas, ou *link prediction*, bem como as abordagens, métricas e ferramentas comumente utilizadas na resolução do problema.

3 Predição de Arestas

Redes sociais são estruturas altamente dinâmicas, que mudam e crescem rapidamente ao longo do tempo através da criação de novas arestas e nodos, que são representantes das interações subjacentes a estrutura social. Entender os mecanismos pelos quais elas evoluem é uma questão fundamental ainda não completamente entendida, definido na literatura como o problema da predição de arestas (*the link prediction problem*). Isto é, dado o estado de uma rede social em um momento t, busca-se prever de maneira assertiva arestas que serão adicionadas a rede durante um intervalo de tempo de t até um momento futuro t' (LIBEN-NOWELL; KLEINBERG, 2004).

O problema de *link prediction* busca responder até que ponto a evolução de uma rede social pode ser modelada utilizando características intrínsecas a própria rede. No sentido de redes de co-autoria, sabe-se que existem inúmeros fatores exógenos a rede que influenciam no porquê de dois autores que nunca colaboraram em um artigo o farão em algum momento futuro.

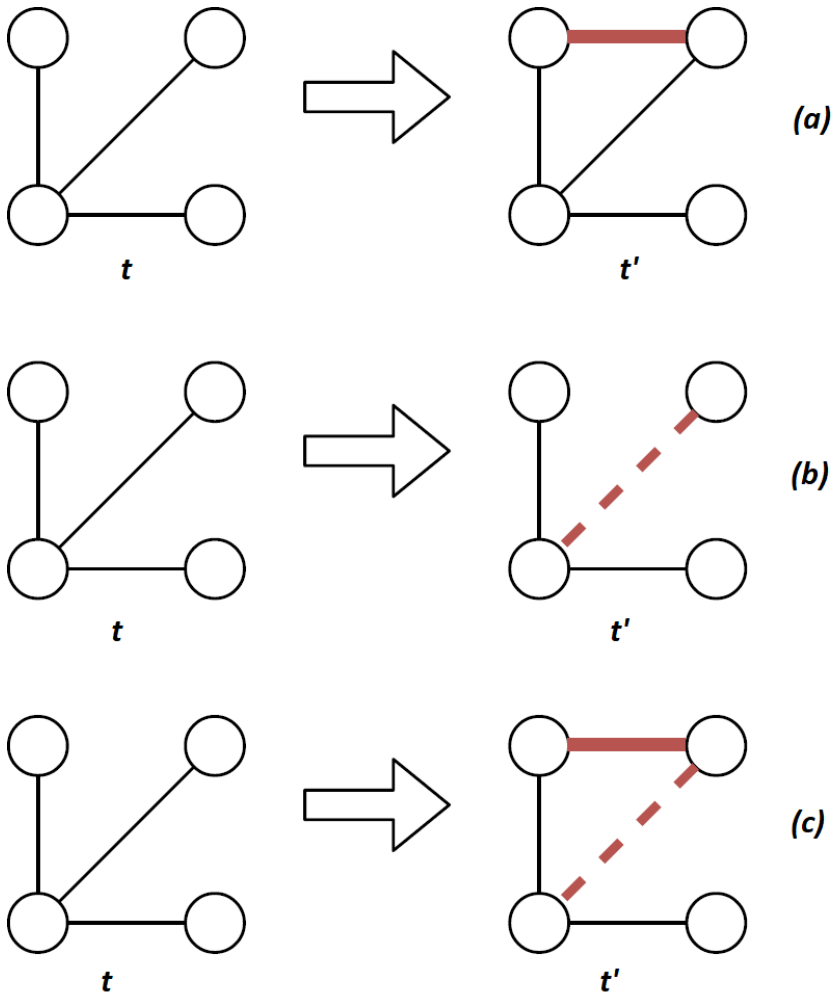
No entanto, alguns destes fatores podem ser sutilmente sugeridos pela estrutura da rede, tais como: colegas pesquisadores em comum, áreas de pesquisa semelhantes, grupos de pesquisa, orientadores e localização geográfica. O problema é observar como tais fatores são representados na rede e utilizá-los em um modelo que possa, de maneira precisa, prever o comportamento da rede. Isso torna palpável a noção intuitiva de semelhança entre indivíduos que auxilia no processo de predição de novas conexões.

A abordagem do problema pode ser vista, em geral, de duas formas: abordagens por medida de similaridade e abordagens com aprendizado de máquina (GAO et al., 2014). A primeira tende a usar medidas de similaridade entre os pares de indivíduos para ranquear os pares e realizar previsões com base em notas de corte no ranqueamento. Ou seja, para cada par de vértices (x, y) não conectados, associa-se uma pontuação de similaridade de acordo com a medida escolhida, onde uma pontuação maior significa uma chance maior de que o par esteja conectado no futuro. Com isso, a lista de pontuações é ordenada de maneira decrescente, e através de uma nota de corte (GAO et al., 2014), ou comparando com os links que realmente ocorreram (LIBEN-NOWELL; KLEINBERG,

2004), os pares previstos são escolhidos.

Já com aprendizado de máquina, em geral, se transforma o problema em classificação binária, onde são escolhidos um ou mais classificadores ou modelos probabilísticos. Neste caso, cada par não-conectado corresponde a uma instância com características (*features*) descrevendo os vértices e o rótulo de classe (*class label*). Este último pode ser rotulado como positivo ou negativo, conforme os pares forem ou não conectados no futuro. Para este tipo de abordagem, as características normalmente são compostas por dois tipos de valores: (i) medidas de similaridade, que correspondem as mesmas utilizadas na primeira abordagem; e (ii) atributos dos vértices, que são provenientes de fontes externas, o que pode incluir questões como localização, área de domínio, informações textuais e outras (HASAN; ZAKI, 2011).

Figura 3.1: Categorias de técnicas e problemas associados a *link prediction*



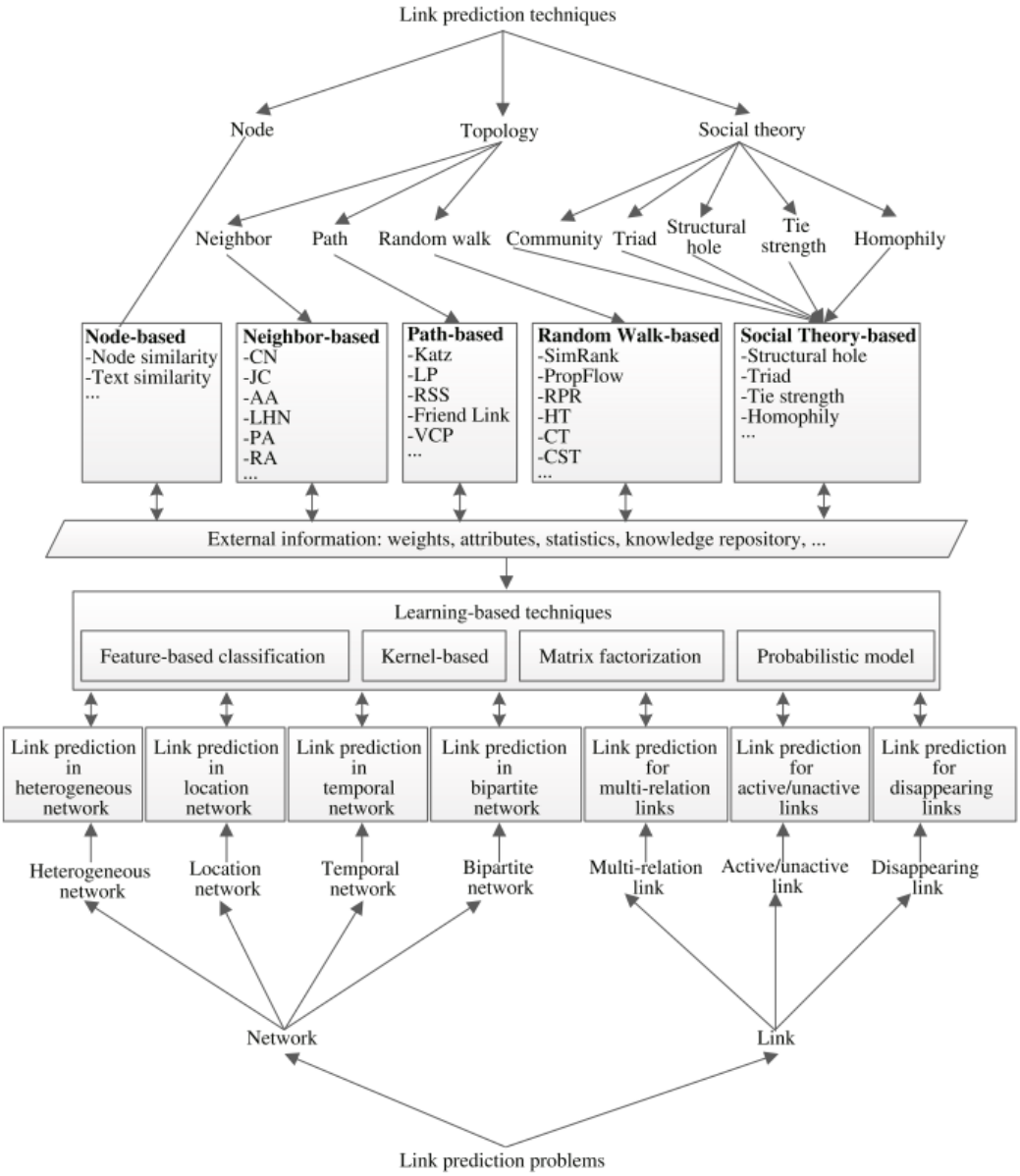
Fonte: O autor, adaptado de (GAO et al., 2014)

O problema tem variantes e aplicações fora do contexto de redes colaborativas,

e de maneira geral, a previsão de arestas especificamente, pode ser dividida em previsão de novas arestas, remoção de arestas, ou ambos, (a), (b) e (c) como pode ser visto na figura 3.1 respectivamente. Considerando o problema descrito no início desta seção, a figura 3.1 mostra estas variantes do ponto de vista de um ponto de tempo t, e um ponto de tempo posterior t', sendo que (a) constitui o problema tradicional. Cada uma destas possibilidades implica em uma abordagem diferente de problema. Para redes de colaboração o problema geralmente é tratado apenas no quesito de novas arestas.

Estes grupos ainda podem ser sub-divididos em mais problemas que não só abrangem redes de colaboração, mas também quaisquer tipos de rede para a qual a predição de arestas pode ser aplicada. Wang et al. (2014) propuseram uma taxonomia que une os tipos de problemas com as técnicas específicas para aborda-los. Os tipos de problemas de previsão de arestas são compostos por duas partes, problemas associados ao tipo da rede, e problemas associados ao tipo de conexão. Isso pode ser visto na parte inferior da figura 3.2.

Figura 3.2: Categorias de técnicas e problemas associados a *link prediction*



Fonte: (WANG et al., 2014)

Os tipos de problema da taxonomia proposta por Wang et al. (2014) são interessantes para notar a abrangência de *link prediction*. Alguns destes são resumidos a seguir:

- **Redes Heterogêneas**: Uma rede heterogênea é caracterizada por múltiplos tipos de arestas, a previsão nestes tipos de rede torna o problema mais complexo, pois técnicas de medidas usadas para previsão em redes homogêneas assumem que os mesmos fatores se aplicam a todas as arestas, logo não podem ser aplicadas a essas redes. Negi e Chaudhury (2016), por exemplo buscaram prever conexões do tipo

usuário-usuário, usuário-grupo e grupo-imagem para um serviço de hospedagem de imagens.

- **Redes Temporais**: o problema para redes temporais é uma extensão do problema tradicional de *link prediction*, onde dadas as informações para o estado do grafo em um tempo T, não só busca-se prever os relacionamentos no tempo $T + 1$, mas também para os tempos $T + 2, T + 3, ..., T + L$, onde L é a largura de um padrão periódico. Essas redes são geralmente vistas em redes de comunicação, como *e-mail* e tráfego de rede, onde interações semanais ou mensais são comuns (DUNLAVY; KOLDA; ACAR, 2010).

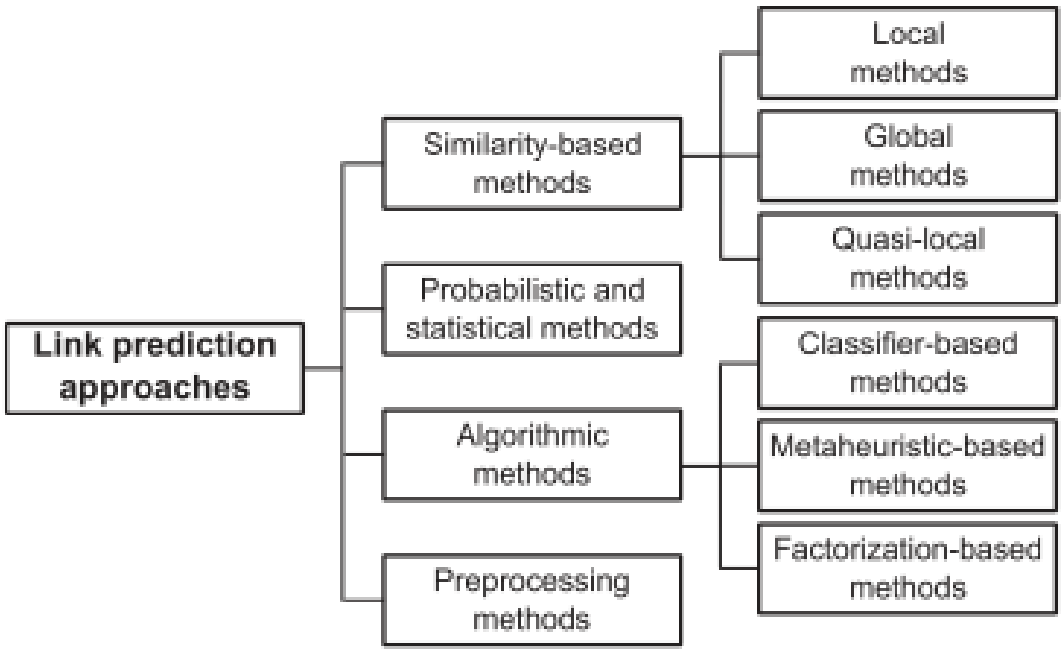
- **Redes Bi-partidas**: como visto anteriormente, redes de co-autoria podem ser representadas por grafos bi-partidos. Redes de usuário-produto em ambientes de *e-commerce* também usam deste tipo de estrutura. Guns (2016) afirma que grafos este tipo carregam mais informações que grafos com arestas ponderadas e não ponderadas, e em seus experimentos conseguiu confirmar que a precisão de abordagens de previsão de arestas podem se beneficiar do uso dessas estruturas.

- **Redes com arestas Ativas/Inativas**: algumas redes sociais também podem avaliar o problema de predição de arestas para conexões que podem estar ou não ativos, como a opção de seguir ou não uma entidade. Como por exemplo a previsão de conexões na rede social *Twitter*, onde entidades mais frequentemente seguem ou deixam de seguir outras entidades de acordo o momento (WANG et al., 2014).

3.1 Técnicas para Predição de Arestas

As técnicas a seguir estão divididas em dois grandes grupos, os métodos de similaridade que estão associados a suas respectivas teorias sociais e apenas podem ser aplicados um por vez, e os métodos de aprendizado, que muitas vezes utilizam de múltiplas medidas e similaridade para gerar uma função que determina a previsão. Para os métodos de aprendizado, vale salientar que a combinação de medidas usadas é importante e afeta a precisão do algoritmo.

Figura 3.3: Taxonomia de Martínez, Berzal e Cubero (2016)

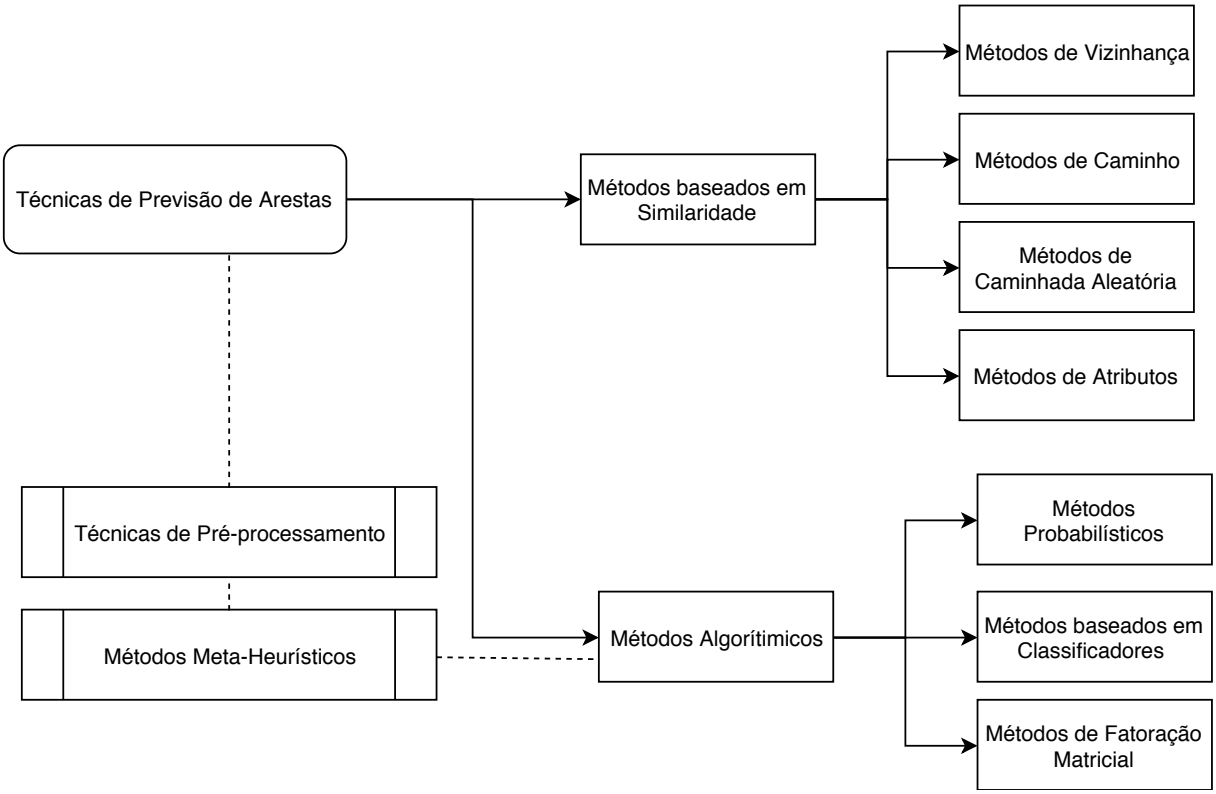


Fonte: (MARTÍNEZ; BERZAL; CUBERO, 2016)

A Figura 3.3, descreve a taxonomia para técnicas de predição de arestas de Martínez, Berzal e Cubero (2016). Esta, foi dividida em quatro abordagens, métodos baseados em similaridade, métodos estatísticos e probabilísticos, métodos algorítmicos, e métodos de preprocessamento. Métodos de similaridade desta taxonomia são análogos aos métodos de nodo, topologia e teoria social do topo da figura 3.2 da taxonomia de Wang et al. (2014). O diferencial é que para o a figura 3.3 a divisão está associada a complexidade de tempo associada as técnicas, enquanto para a figura 3.2 a divisão está associada a estratégia usada para a geração da medida de similaridade.

A taxonomia descrita na Figura 3.4 é a usada neste trabalho, e constitui uma combinação entre as propostas de Wang et al. (2014) e Martínez, Berzal e Cubero (2016). Nela, as medidas de similaridade são baseadas em Wang et al. (2014) e as técnicas algorítmicas de acordo com Martínez, Berzal e Cubero (2016). A intenção desta combinação foi tornar homogênea e clarificar a hierarquia dos métodos.

Figura 3.4: Taxonomia Proposta sobre técnicas para predição de arestas



Fonte: O autor

Na Figura 3.4, pode-se observar algumas mudanças no modelo de Martínez, Berzal e Cubero (2016), como por exemplo, separar métodos meta-heurísticos e de pré-processamento. Esta separação ocorre pois o uso dessas abordagens é opcional e não resulta diretamente em uma previsão. Outra mudança foi mover métodos probabilísticos para dentro de métodos algorítmicos. Essa escolha foi feita considerando a definição da palavra algorítimo, como: "uma sequência finita de instruções bem definidas e não ambíguas, cada uma das quais devendo ser executadas mecânica ou eletronicamente em um intervalo de tempo finito e com uma quantidade de esforço finita"(MEDINA, 2005), o que acaba por englobar estes métodos.

Isso resulta em uma taxonomia que separa as técnicas para predição de arestas em dois grupos, Métodos Baseados em Similaridade e Métodos Algorítmicos. O primeiro é composto por métodos baseados em vizinhança, caminho, caminha aleatória, e atributos, os quais estão associados às teorias sociais da seção 2.3. O segundo é composto por métodos probabilísticos, que estão associados especificamente a estatística e matemática, métodos baseados em classificadores, os quais usam de aprendizado de máquina supervisionado, e métodos de fatoração matricial, os quais usam de filtragem colaborativa, vistos

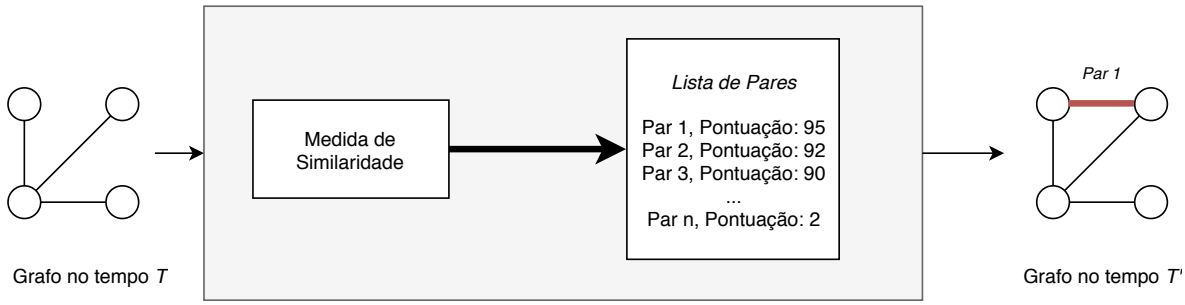
principalmente em sistemas de recomendação.

As subseções a seguir buscam deixar mais claras as formas de abordagem para estas técnicas, com enfoque em métodos baseados em medidas de similaridade e a subclasse de métodos algorítmicos, métodos baseados em classificadores. Isso se dá devido a um maior apoio da literatura com viés computacional a estas técnicas.

3.1.1 Métodos Baseados em Similaridade

Métodos baseados em similaridade assumem que arestas tendem a se formar entre vértices semelhantes. Estes métodos seguem a hipótese que dois vértices são similares se eles se comportam de maneira parecida na rede, com vizinhos em comum ou próximos entre si dada uma função de distância. De maneira geral, essas abordagens definem uma função $s(x, y)$, que associa uma pontuação de semelhança, para um par de nodos x e y. Esta função é computada para todos os pares de vértices relevantes, normalmente os sem conexão direta entre si. A partir disso, lista-se em ordem decrescente os pares com pontuações de similaridade mais altas, sendo estes os mais prováveis para uma conexão futura, como ilustra a Figura 3.5.

Figura 3.5: Estrutura padrão para predição baseada em similaridade



Fonte: O autor

Assim, estes refletem na informação sobre a topologia da rede. O progresso nessa área se dá principalmente devido a estudos de matemáticos e físicos. Vários autores realizaram levantamentos e descrevem medidas de similaridade como Hasan e Zaki (2011), Linyuan e Zhou (2011), Gao et al. (2014) e Wang et al. (2014), sendo que as medidas a seguir são derivadas destes trabalhos. Dado o número vasto de medidas utilizadas na área, são citadas apenas algumas mais frequentemente aplicadas em trabalhos relacionados.

Alguns autores como Linyuan e Zhou (2011) e Martínez, Berzal e Cubero

(2016) dividem os métodos baseados em similaridade em Locais, Globais, e Quasi-locais (figura 3.3), que fazem parte de uma taxonomia principalmente baseada em complexidade de tempo. Neste estudo, a divisão é feita por tipos de medidas em questões de proximidade, e por consequência, de acordo com suas respectivas teorias sociais, sendo os métodos de: Vizinhança, Caminho, Caminhada Aleatória, e Atributos, semelhante a taxonomia proposta por Wang et al. (2014).

Nota-se que os símbolos x e y representam nodos, $\Gamma(x)$ e $\Gamma(y)$ denotam conjuntos de vizinhos destes nodos, e k_x e k_y o grau dos mesmos.

Métodos de Vizinhança

Essas abordagens usam informações relacionadas a vizinhança para computar a similaridade. A maior desvantagem destes métodos é que eles estão limitados a avaliação de nodos de no máximo distância 2, ou seja, vizinhança imediata ou direta (MARTÍNEZ; BERZAL; CUBERO, 2016). Mesmo com esta limitação e a relativa simplicidade, estes métodos ainda são bastante válidos mesmo quando comparados a estratégias mais complexas. Destes os métodos mais conhecidos são:

- ***Common Neighbours - CN:*** o método de vizinhos em comum, em português, baseia-se na premissa de que dois indivíduos que dividem um número considerável de conhecidos, mas não se conhecem tem chances maiores de o fazer no futuro. Esta abordagem é as vezes utilizada como base para julgar a performance de outros métodos (LIBEN-NOWELL; KLEINBERG, 2003), devido a sua simplicidade. A medida é definida por:

$$s(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

- ***Coeficiente de Jaccard - JC:*** O coeficiente de Jaccard ou índice de Jaccard é baseado no método de vizinhos em comum, e é uma medida estatística para comparar a similaridade de conjuntos. Em *link prediction*, a ideia é comparar a similaridade dos conjuntos de vizinhos, os normalizando para acrescentar proporcionalidade. Esta medida é definida por:

$$s(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

- ***Índice de Adamic/Adar - AA:*** esta medida teve como propósito inicial computar a similaridade entre páginas pessoais (ADAMIC; ADAR, 2003), e assume que em uma rede social real, se um conhecido comum entre duas pessoas tem mais amigos, então é menos provável que ele apresente essas duas pessoas, ou no caso contrário, se ele conheça menos pessoas a chance de que ele as apresente é maior. O índice é definido por, onde z é um nodo vizinho de x e y:

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

- ***Preferential Attachment - PA:*** este método segue a ideia de que nodos com maior grau tendem a se conectar a mais nodos. A medida surgiu como resultado direto do estudo de Barabási e Albert (1999), vinculado a redes sem escala. Este estudo chegou a conclusão de que devido a razão proporcional a lei de potência da distribuição de grau em redes complexas, há maior probabilidade de um nó com um número elevado de vizinhos se conectar a mais nós ao longo do tempo. Este método também tende a ser utilizado como método para comparação com outros métodos. Ele visa calcular a semelhança para cada par de nodos, não só os conjuntos de vizinhos. Pode ser definido por:

$$s(x,y) = |\Gamma(x)| * |\Gamma(y)|$$

- ***Resource Allocation - RA:*** este índice é motivado pelo processo de alocação de recursos que ocorre em redes complexas. Ele modela a transmissão de recursos entre dois nodos através da vizinhança. Cada nodo vizinho recebe uma unidade de x, e igualmente a distribui a seus vizinhos, os recursos que vão a direções opostas tendem a se dissipar e a quantidade de recursos obtidos por y é usada como a similaridade entre este e x(LÜ; JIN; ZHOU, 2009).

$$s(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|}$$

Outras medidas incluem: *Índice de Salton*, também conhecido como similaridade de cossenos; *Índice de Sørencen*, usado para comparar bases de dados ecológicas; *Hub-Promoted Index* e *Hub-Depressed Index*, usados em redes metabólicas; e *Índice de*

Leicht-Holme-Newman, um variante mais sensível do coeficiente de Jaccard.

Métodos de Caminho

Medidas baseadas em caminho, estão associadas a teoria do balanço cognitivo, e usam um ou mais caminhos entre os vértices x e y para computar a similaridade. São mais abrangentes que as medidas de vizinhança, pois podem considerar até os vértices mais distantes entre si na rede.

- ***Caminho Mínimo - CM:*** A medida de caminho mínimo é a medida mais simples deste tópico, e pela sua simplicidade as vezes é usada como medida comparativa (LIBEN-NOWELL; KLEINBERG, 2004). Basicamente, dados dois vértices em um grafo, o caminho mínimo é a menor distância entre estes, em termos de menor número de arestas percorridas de x a y.

- ***Índice de Katz:*** Proposto por Katz (1953), este índice busca somar todos os caminhos possíveis entre dois vértices, e exponencialmente penaliza-los pelo seu tamanho.

$$\sum_{l=1}^{\infty} \beta^l \cdot |paths_{xy}^{<l>}|$$

  Onde $paths_{xy}^{<l>}$, são os caminhos de tamanho l entre x e y, e β, um fator de ajuste que quanto maior, dá mais importância à caminhos mais longos ($0 < \beta < 1$).

- ***Local Path Index - LPI:*** este método é derivado o índice Katz, mas apenas considera um número limitado de caminhos, o que o torna mais viável computacionalmente (LÜ; JIN; ZHOU, 2009).

- ***FriendLink - FL:*** A ideia dessa medida é que pessoas em uma rede social podem usar todos os caminhos disponíveis para formar uma conexão, sendo que caminhos mais curtos são mais importantes que mais longos. Ela é outra variante da medida Katz, porém propõe um fator de atenuação de caminhos mais longos diferente (PAPADIMITRIOU; SYMEONIDIS; MANOLOPOULOS, 2012).

Métodos de Caminhada Aleatória

Interações sociais também podem ser modeladas através de caminhadas aleatórias, que usa múltiplas transições com probabilidade de um nodo para seus vizinhos para denotar similaridade. Estes métodos são associados a teoria de balanço cognitivo.

- ***Random Walk - RW:*** este método assume que, se a partir de um grafo e um nodo inicial, seja selecionado aleatoriamente um vizinho deste, iniciando um caminho, e repetir este processo para cada vértice alcançado temos uma caminhada aleatória. Diversas medidas são derivadas desta para a predição de arestas, incluindo sua variante *Random Walks with Restart* (TONG; FALOUTSOS; PAN, 2006). A medida de similaridade é definida pela probabilidade média aproximada de x alcançar y em uma caminhada. Inicialmente proposta por Pearson (1905), para descrever processos estocásticos em áreas como economia, biologia e física.

- ***Propflow:*** esta medida é proporcional a probabilidade de que um *random walker* começando em x alcance y, no entanto a caminhada é restrita, ela seleciona arestas baseadas em pesos e termina quando re-visita vértices ou chega no destino (y). Isso serve como estimativa para prever novas conexões (LICHTENWALTER; LUSSIER; CHAWLA, 2010).

- ***SimRank - SR:*** a similaridade desta medida busca avaliar quanto tempo dois *random walkers*, que saem de x e y demoram para se encontrar. Esta medida é bastante custosa computacionalmente, com complexidade $O(n^4)$ (JEH; WIDOM, 2002).

- ***Local Random Walks - LRW:*** nesta medida, usa-se o conceito de caminhadas aleatórias, porém limita o número de iterações a um número fixo. Isso torna a medida mais viável em termos de complexidade (LIU; Lü, 2010).

Métodos de Atributos

Os métodos baseados em atributos, ou informações nodais, são baseados na teoria de homofilia e seguem a simples premissa de que quanto maior a semelhança entre os vértices, maior a probabilidade de eles se conectarem. As informações usadas para estes métodos são externas, ou seja, não estão visualmente representadas no grafo, e podem incluir:

dados de filiação, áreas de interesse, perfis sociais, palavras-chave em artigos, número de publicações em comum e várias outras (WANG et al., 2014), (HASAN; SALEM; ZAKI, 2006).

A desvantagem deste tipo de medida é que não são todas as situações que este tipo de informação está disponível. E usualmente, estas informações são usadas em conjunto a outras medidas de similaridade estrutural em abordagens que utilizam de aprendizado de máquina, com o intuito de melhorar a performance da previsão (HASAN; ZAKI, 2011).

3.1.2 Métodos Algorítmicos

Métodos algorítmicos usam de abordagens computacionais para a resolução do problema. Estes métodos tendem a ser mais robustos por conseguirem utilizar de múltiplas medidas de similaridade a seu favor. Nesta subseção dar-se-a ênfase aos métodos baseados em classificadores, pois são os mais estudados na literatura (HASAN; ZAKI, 2011),(MARTÍNEZ; BERZAL; CUBERO, 2016), e possuem o maior número de ferramentas disponíveis para desenvolvimento e avaliação. Isto é verificado com mais detalhes no Capítulo 4.

Métodos baseados em Classificadores

Estes métodos utilizam de aprendizado de máquina supervisionado (*supervised machine learning*), que busca encontrar, ou aprender, uma função que mapeia uma entrada de dados a uma saída, ou resposta, através do uso de pares de entrada-saída já mapeados. O conjunto de pares entrada-saída já mapeados são chamados de conjunto de treinamento (*labeled training data*). Estes pares consistem, normalmente, de vetores de características (*features*), para a entrada, e a saída pode ser um valor numérico objetivo ou uma etiqueta de classe (*class label*). O objetivo da aprendizagem supervisionada é criar uma função capaz de prever o valor correspondente a qualquer entrada válida depois de ter visto vários exemplos, os dados de treinamento. Para isso, o algoritmo busca generalizar as escolhas a partir dos dados apresentados para situações não vistas (HASAN; ZAKI, 2011).

Para a resolução do problema de predição em redes de co-autoria, em geral, considera-se o problema em classificação binária. Dado um grafo e seus vértices, cada par de vértices não conectados correspondem a uma instância da classe, podendo esta ser

marcada como positiva (há conexão) ou negativa (não há conexão) de acordo a previsão do classificador (WANG et al., 2014).

Um classificador define como o sistema caracteriza e avalia os dados a serem previstos, ou sem etiquetas (*unlabeled data*). Não existe uma forma definitiva para todas a bases ou problemas, e uma variedade de algoritmos. Kotsiantis (2007) realizou um levantamento destes. Vários podem ser aplicados ou adaptados para o problema classificação binária, e alguns destes são descritos brevemente a seguir:

- ***Naive Bayes NB:*** essa técnica é baseada no teorema de Bayes, que descreve a probabilidade de um evento ocorrer com base em conhecimento anterior sobre os mesmo. Este classificador assume que as características usadas são independentes, por isso o termo Naive.

- ***Support Vector Machines - SVM:*** Este método transforma o número de características (N) usadas para a previsão em um hiperespaço N-dimensional, e busca encontrar o hiperplano que melhor separa as classes no mesmo, ou seja, o hiperplano que possui a maior margem de distância entre as classes de dados.

- ***Árvores de Decisão - DT:*** O aprendizado nesta abordagem consiste na construção dessa árvore a partir dos dados de treinamento. Esta estrutura usa de observações de um item, representado nos ramos da árvore, para chegar a conclusões sobre o valor objetivo, representado nas folhas da árvore. Existem métodos agregados (*ensemble methods*) que constroem mais de uma árvore, como agregação por *bootstrap* e potenciação de gradiente (*boosted trees*).

- ***Regressão Logística:*** classificação por regressão logística usa de uma equação linear que combina as características usadas para realizar previsões, e utiliza de uma função sigmoide para alocar as probabilidades das previsões em valores entre 0 e 1.

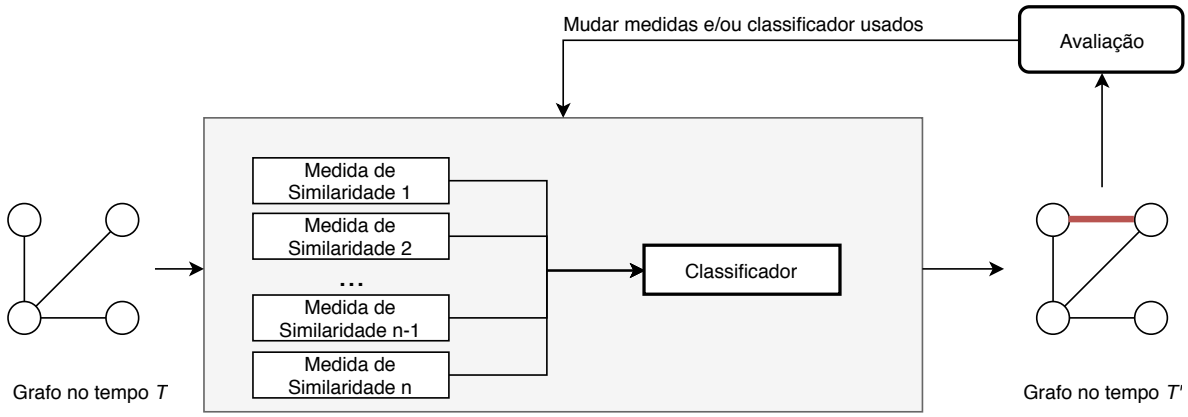
- ***Redes Neurais - RN:*** redes neurais em aprendizado de máquina buscam simular a estrutura neural humana, processando registros um por um e comparando sua classificação inicial com os valores reais e aprendendo a partir dos erros ou acertos.

Para construir um classificador eficiente é necessário definir e extrair um conjunto de características apropriadas da rede. Neste caso, as características podem ser

medidas de similaridade topológicas ou específicas do domínio, ou seja, informações que podem ser obtidas apenas observando a rede ou informações externas.

Um dos benefícios diretos do uso deste tipo de abordagem é a possibilidade de utilizar mais de uma característica para realizar a previsão. O que permite o estudo de que fatores influenciam ou influenciaram a evolução da rede durante o período analisado (MARTÍNEZ; BERZAL; CUBERO, 2016). Não apenas isso, mas as previsões também estão associadas ao tipo de classificador usado, portanto a avaliação de combinações modelos e características diferentes é interessante para obter melhores resultados. A Figura 3.6, mostra de maneira generalizada o procedimento de previsão de arestas baseada em classificadores.

Figura 3.6: Estrutura padrão simplificada para predição baseada em classificadores



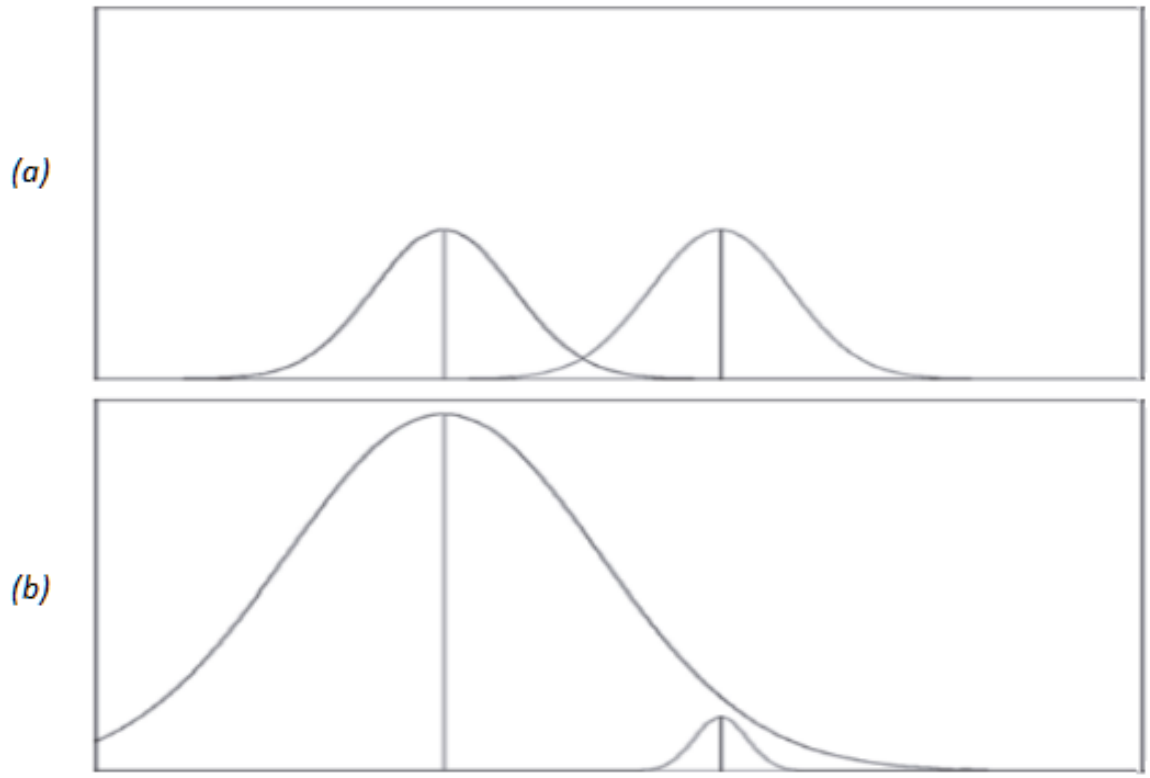
Fonte: O autor

No entanto, estes métodos também trazem as suas dificuldades. Como por exemplo, o extremo desbalanceamento de classe, que em *machine learning* é um problema recorrente, onde ocorre que o número total de uma classe é muito maior que outra. Isso ocorre em *link prediction* pois, em geral, o número de arestas possíveis é quadrático ao número de vértices no grafo, mas o número de arestas reais é apenas uma pequena fração deste número. Por consequência, classificadores têm dificuldade de prever a existência ou não de uma aresta no futuro, visto que na grande maioria dos casos não há arestas entre dois vértices, tende-se a prever que não há arestas sempre (HASAN; ZAKI, 2011).

Por exemplo, se dados 100 pares de autores e na realidade apenas 20 desses pares realizaram co-autorias para o espaço de tempo em questão, e o classificador assumisse que todas as instâncias são negativas, ainda teria-se acurácia de 80%. Enquanto um classificador que previsse 10 dessas co-autorias, mas também previsse outras 30 que

não ocorreram, teria uma acurácia de 60%. Para a recomendação de arestas, o objetivo é prever novas conexões, que o primeiro modelo não identifica, mas o segundo apesar de errar, consegue recomendar corretamente alguma. Portanto, neste caso o segundo classificador é melhor, apesar da métrica de avaliação não refletir essa questão. Isso relembra a importância da escolha das métricas de avaliação corretas, uma vez que quando considerada apenas acurácia, os resultados acabam sendo tendenciosos. Para de medidas de precisão refira a seção 3.2.

Figura 3.7: (a), representação da distribuição de classes balanceada. (b), a mesma representação para uma distribuição de classes desbalanceada



Fonte: (HASAN; ZAKI, 2011)

A Figura 3.7, demonstra o que o algoritmo de classificação consegue extrair a partir das características para os casos de distribuições de classe balanceada e desbalanceada. Observa-se que em (a), os montes estão distintos, o que facilita a separação dos mesmos. Por outro lado, em (b), observa-se o efeito do desbalanceamento de classe, onde classes em minoria tendem a ser sub-representadas a favor da classe majoritária.

Métodos Probabilísticos

Modelos de redes podem ser descritos de acordo com princípios estatísticos e probalísticos. Estes métodos assumem que a rede tenha uma estrutura conhecida, é construído um modelo específico para esta, e estimam os parâmetros usando técnicas estatísticas. Os parâmetros são usados para computar a probabilidade de formação de novas arestas, que são ranqueadas de maneira semelhante a métodos baseados em similaridade. Alguns exemplos são: modelo de estrutura hierárquica, modelo de bloco estocástico e modelo de formação de ciclos (MARTÍNEZ; BERZAL; CUBERO, 2016).

Métodos baseados em Fatoração Matricial

Métodos de fatoração matricial são largamente utilizados em sistemas de recomendação, que podem fazer uso de características latentes ou características adicionais para realizar a predição. Esses métodos fazem parte da família de técnicas de filtragem colaborativa, que tem a premissa de que se uma entidade A e uma entidade B concordam com algo, A têm mais chances de concordar com B em outro assunto, do que uma entidade qualquer. Huang, Li e Chen (2005) estudam o problema para o caso de predição de arestas, e obtiveram resultados inicias favoráveis.

Métodos Meta-heurísticos

Métodos meta-heurísticos são procedimentos de alto-nível desenvolvidos para encontrar, gerar ou selecionar outras heurísticas que possam providenciar uma solução suficientemente boa para o problema. Em geral estas abordagens são menos exploradas. Um exemplo de trabalho que usou disso foi Bliss et al. (2013), que desenvolveu um algoritmo evolutivo que assume que diferentes medidas de similaridade podem co-existir e co-operar na mesma rede. Neste caso, o algoritmo usa a estratégia evolutiva para otimizar a influência das medidas heurísticas entre si.

3.1.3 Técnicas de Pré-processamento

Métodos de pré-processamento também são conhecidos como abordagens de alto nível ou meta-abordagens, visto que a intenção é utilizar essas técnicas em conjunto com outros

métodos. Seu maior objetivo é reduzir arestas "fracas"ou "falsas", visando melhorar a performance do método aplicado para predição.

O método mais simples destes é Filtragem, que consiste em remover pares da avaliação antes da aplicação do algoritmo, de acordo com um ou mais atributos ou medidas de similaridade que obtêm resultados de semelhança muito baixos. Bigramas não-vistos (*Unseen Bi-grams*), que é geralmente aplicada na área linguística, usa de semelhança entre vértices para inferir ou substituir outros vértices próximos, mas também pode ser adaptada para as heurísticas mais comuns, visando melhorar a performance por não ter de avaliar todos os vértices (LIBEN-NOWELL, 2005). Outros métodos incluem aproximação matricial, que busca reduzir o número de nós convertendo o grafo em um problema de otimização que visa minimizar o número de vértices semelhantes (MARTÍNEZ; BERZAL; CUBERO, 2016).

3.2 Métricas de Avaliação

Para a avaliação de um modelo preditivo são necessários um número de métricas que clarifiquem o resultado, e são essenciais pra avaliação do classificador ou método de modo comparativo (HOSSIN; SULAIMAN, 2015). Para que elas possam ser descritas, considera-se que em um problema de classificação existem duas possibilidades de saída: positiva (p) ou negativa (n), e que podem ser caracterizadas de quatro maneiras: se a saída do modelo for p e a resposta for p tem-se um verdadeiro positivo (VP), caso a resposta seja n então ele é considerado um falso positivo (FP). Por outro lado, se a predição for n e o valor real seja n tem-se um verdadeiro negativo (VN), porém se o valor real for p então há um falso negativo (FN). P e F correspondem ao número de instâncias positivas e negativas, respectivamente. Com isso pode-se calcular uma série de métricas (POWERS, 2007):

Precisão

Precisão é definida como a proporção entre os verdadeiros positivos e todas as predições positivas.

$$Precisão = \frac{VP}{VP + FP}$$

Revocação

Revocação (ou *Recall*), também conhecida como sensibilidade, é a proporção entre verdadeiros positivos, e o total de verdadeiros real, ou seja a soma entre o número de verdadeiros positivos e falsos negativos.

$$Revocação = \frac{VP}{VP + FN}$$

Medida F

A medida F, valor F ou *f1-score*, é a média harmônica de precisão e revocação. Um valor de F mais próximo de 1 é melhor, enquanto um mais próximo de 0 é pior.

$$F1 = \frac{Precisão \cdot Revocação}{Precisão + Revocação}$$

Acurácia

Acurácia é definida como a proporção entre acertos (verdadeiros positivos e verdadeiros negativos) e o total de casos avaliados.

$$Acurácia = \frac{VP + VN}{P + N}$$

AUC

AUC (*Area Under Curve*) ou área abaixo a curva, mais especificamente abaixo a uma curva COR (Característica de Operação do Receptor), é uma representação gráfica usada para mostrar o desempenho de um classificador binário. Esta medida é obtida pela razão entre a Revocação e a razão de falsos positivos (RFP):

$$RFP = \frac{FP}{FP + VN}$$

Curvas COR tem vantagens quando comparadas a curvas de precisão e revocação, e por isso é usada para medir a qualidade de classificadores (FAWCETT, 2004). Elas podem ser usadas para avaliar a habilidade do modelo de discriminar pares de autores

que tem novas colaborações e aqueles que não. Por outro lado, Yang, Lichtenwalter e Chawla (2015) argumentam que para o caso de predição de arestas a medida não reflete completamente a verdade, pois devido ao extremo desbalanceamento de classe, curvas AUC podem mascarar os resultados.

3.3 Ferramentas

Para a resolução e avaliação do problema de modo mais eficiente, autores empregam diversas ferramentas que facilitam as tarefas. Algumas destas são listadas a seguir:

- ***Weka Project***: Essa ferramenta foi desenvolvida pela Universidade de Wakaito, na Nova Zelância, e é uma plataforma *open-source* implementada em Java, sobre a Licença Pública Geral GNU (*GNU General Public License*), que dispõem de uma variedade de algoritmos de aprendizado de máquina. A plataforma visa tornar técnicas de *Machine Learning* disponíveis para todos (HALL et al., 2009).

- ***Python scikit-learn***: é uma biblioteca para a linguagem Python, construída com base nas bibliotecas SciPy, NumPy e matplotlib, *open-source* pela licença BSD que disponibiliza ferramentas eficientes para mineração e análise de dados. A ferramenta dispõe de módulos para classificação, regressão, agrupamento, redução e dimensionalidade, seleção de modelos e pré-processamento de dados (PEDREGOSA et al., 2011).

- ***Python linkpred***: é uma biblioteca Python para predição de arestas, que providencia um número de heurísticas (técnicas com base em similaridade). Esta sob a licença BSD, logo é *open-source* e pode ser usada por linha de comando (GUNS, 2016).

- ***Python networkX***: é uma biblioteca Python para criação, manipulação, e estudo da estrutura, dinâmica e funções de redes complexas. Nela estão implementados algoritmos padrão para grafos, geração de estruturas clássicas, aleatórias e mais. É mais uma ferramenta *Open-source* pela licença BSD (HAGBERG; SCHULT; SWART, 2008).

- ***MATLAB***: um ambiente para programação multi-plataforma e linguagem de programação proprietária desenvolvida por *MathWorks*, e pode interagir com diversas

outras linguagens de programação incluindo C/C++, Python, Fortran e Java. Esta foi usada por Hasan, Salem e Zaki (2006).

- ***SPSS Clementine***: é uma ferramenta proprietária para mineração de dados e análise de texto desenvolvida pela IBM, que também pode ser usada para construir modelos preditivos e outras tarefas analíticas. Utilizada em Bartal, Sasson e Ravid (2009)

- ***Gephi***: é uma ferramenta *open-source* para visualização e exploração de grafos, que permite análise de conexões, análise de redes sociais, análise exploratória de dados e mais. Possui uma série de métricas de avaliação já implementadas e está disponível para Linux, Windows e Mac OS (BASTIAN; HEYMANN; JACOMY, 2009).

3.4 Considerações sobre o Capítulo

Neste capítulo foi apresentada a importância da recomendação ou previsão de novas conexões para redes complexas, bem como os problemas e técnicas associados. De acordo com o quesito técnicas, foram destacados dois tipos: abordagens que usam apenas medidas de similaridade e abordagens que aplicam classificadores. A vantagem principal do uso de apenas medidas de similaridade está na facilidade de computação e avaliação, porém a desvantagem dessas medidas é que as previsões são realizadas com base na medida isoladamente, sem o o auxílio de outras informações. Já os classificadores tem como principal vantagem o uso de múltiplas características de similaridade simultaneamente, todavia possuem desvantagens como a questão do desbalanceamento de classe que podem comprometer os resultados. Por fim, foram apresentadas métricas para a avaliação dos modelos de predição e ferramentas comumente utilizadas na área.

Os conceitos apresentados neste capítulo são fundamentais para o entendimento da área, ainda crescente, de predição de arestas. No Capítulo 4, são avaliados alguns trabalhos que tratam do problema especificamente para o caso de redes de co-autoria. As noções de medidas de similaridade, seu uso em conjunto ou não a classificadores e a respectiva avaliação das previsões de modelos dos trabalhos relacionados são analisados pelo próximo capítulo.

4 Trabalhos Relacionados

Este capítulo está dividido em duas partes: a primeira dedica-se aos artigos que aplicam experimentos de predição de arestas através do uso de medidas de similaridade, e que não aplicam técnicas baseadas em classificadores; a segunda se dedica a trabalhos que utilizam de classificadores para a resolução do problema.

Os artigos apresentados neste capítulo e nos capítulos anteriores, foram obtidos por buscas nas bibliotecas digitais ACM e IEEE, com a palavra-chave "*link prediction*", com 372 e 510 artigos respectivamente. Estes artigos artigos foram filtrados de acordo com a relevância para este trabalho, desde que pudessem ser associados a uma de três categorias, estes seriam inclusos: (a) aplicação de técnicas de previsão de arestas, (b) estudos sobre previsão de arestas e (c) levantamentos ou *surveys* sobre previsão de arestas. Dos artigos obtidos, 27 foram caracterizados com estudos, 15 como aplicação de técnicas, e 6 levantamentos ou *surveys*. Deste último tipo têm-se: Linyuan e Zhou (2011), Hasan e Zaki (2011), Gao et al. (2014), Wang et al. (2014) , Li, Fang e Sheng (2015), Yang, Lichtenwalter e Chawla (2015) e Martínez, Berzal e Cubero (2016), que junto aos estudos, foram os artigos nos quais a fundamentação teórica deste trabalho está embasada.

Dos 15 artigos de aplicação de técnicas, que é tipo de artigo a ser discutido neste capítulo, apenas 9 descreveram a metodologia computacional por trás do trabalho. Por isso apenas estes são discutidos a seguir.

4.1 Trabalhos baseados em Medidas de Similaridade

As subseções seguintes referem-se aos trabalhos baseados em medidas de similaridade, ou seja que utilizam de medidas de similaridade isoladamente. Estes métodos são discutidos na seção 3.1.1, e podem ser divididos em medidas de similaridade de vizinhança, caminho, caminhada aleatória e atributos.

4.1.1 Linben-Nowell e Kleinberg (2003)

O estudo realizado por Liben-Nowell e Kleinberg (2003) definiu formalmente o problema de predição de arestas, e abriu o caminho para novos trabalhos na área. Em seu estudo, além da apresentação de algumas medidas de similaridade, estas também foram computadas entre pares de pesquisadores em cinco redes de co-autoria para realizar previsões. As redes, por sua vez, foram extraídas de seções diferentes da biblioteca digital *arXiv*. Para cada uma delas, foi criada uma lista de pontuação de similaridade de acordo com a medida, onde arestas previstas como novas formavam um conjunto, o qual seria avaliado com relação a quais arestas da fato se formaram ou não. As medidas usadas foram: Vizinhos em Comum; Conexão Preferencial; Coeficiente de Jaccard; Coeficiente de Adamic/Adar; *SimRank*; *hitting time*; *PageRank*; e Katz.

Como base de comparação foram utilizadas a medida de Vizinhos em Comum, a de Caminho mínimo e o método de previsão aleatória. Através disso, pôde-se observar que essas medidas são consideravelmente melhores que um previsor aleatório, que variou de $0,1\%$ a $0,5\%$ entre as bases, enquanto Adamic/Adar e Katz encontravam uma taxa de acerto que variava de 15% até 55%. O método *hitting time* obteve os piores resultados, sendo o único cuja taxa de acerto não passou do preditor de Caminho Mínimo. Seu estudo, não só mostrou que medidas de similaridade podem ser importantes para o problema de predição de arestas, mas também mostrou a grande variabilidade entre os resultados das diversas redes, o que implica que a disposição da rede pode afetar de maneira significativa as previsões, e com isso a performance do tipo de medida adotada. Por exemplo, observou-se que para uma das bases o método de Conexão Preferencial obteve os melhores resultados, porém para o resto das redes seus resultados não eram expressivos.

4.1.2 Brandão et al. (2013)

Em seu trabalho Brandão et al. (2013) introduzem medidas baseadas em Análise de Redes Sociais (ARS) para a recomendação ou intensificação de colaborações em redes de co-autoria. São propostas medidas de similaridade com base em afiliação e localização geográfica, que seguem os princípios de homofilia e proximidade associados a ARS. As rede de co-autoria avaliadas no trabalho são do CiênciaBrasil e DBLP, com 340 pesquisadores e 629 pesquisadores, da área de computação respectivamente. Os anos avaliados são para CienciaBrasil 2000-2009 para treino, e 2009-2011 para testes, enquanto para DBLP,

1971-2011 para treino e 2011-2012 para testes.

A primeira métrica introduzida $Affin$, segue o princípio da homofilia, é considerada para um par de pesquisadores i, j o número de publicações que i fez com pesquisadores da mesma afiliação institucional de j, dividido pelo número de publicações de i. A segunda métrica é GLI (*Geographic Location Information*), segue o princípio da proximidade, e é basicamente a distância geográfica entre as instituições de dois autores i, j. O modelo criado é avaliado em questões de acurácia, novidade, diversidade e cobertura. Os autores afirmam que a medida de novidade proposta está associada ao quanto o algoritmo trás conexões novas, ou seja que não seriam visíveis sem o algoritmo. As duas últimas, estão geralmente associadas a sistemas de recomendação. Sendo que diversidade está associada a similaridade entre os autores recomendados para colaboração, e a cobertura computar a diferença entre colaborações recomendadas, medindo assim a abrangência do modelo.

No experimento observou-se alto correlacionamento entre colaborações concretizadas e a métrica $Affin$, além de boa acurácia, acima de 80%. Já GLI não obteve resultados satisfatórios em questões de acurácia, mostrando que colaborações estão pouco relacionadas a distância geográfica. No entanto, as recomendações desta medida foram as que propuseram maior novidade e diversidade.

4.1.3 Gao et al. (2014)

Gao et al. (2014) observam o comportamento de dez métodos de previsão baseados em similaridade, com o apoio de diferentes métricas para a avaliação de redes. Sendo eles: Vizinhos em Comum, índice de Jaccard, *Preferential Attachment*, Adamic/Adar, similaridade de cossenos, Katz, Sørensen, *Hub Promoted Index*, *Hub Depressed Index* e índice de Leitch Holme Newman.

Neste trabalho foram utilizadas bases de dados de seis redes sociais demarcadas por intervalos de tempo: Enron, *Facebook*, *Flickr*, PWr , UC Urvine , e *Youtube*. Foi observado que todas as redes seguem uma distribuição de grau semelhante a lei de potência, logo todas são redes livres de escala. Também foram determinadas redes favoráveis a previsões e redes não-favoráveis a previsões. Para as redes favoráveis notou-se que estas tendem a ter coeficientes de agrupamento local e global maiores, bem como um caminho mínimo médio pequeno, características comumente associadas a redes de pequeno mundo.

Tal fato sugere que este tipo de rede é mais tranquilamente prevista através destes modelos quando comparadas a redes aleatórias. Em seus testes observou-se que Katz obteve a maior performance geral. Enquanto Preferential Attachment foi a medida que obteve performance mais consistente, ou seja os resultados dessa medida apesar de não serem os melhores, obtêm resultados aceitáveis para todas as bases.

4.2 Trabalhos baseados em Classificadores

As subseções a seguir descrevem trabalhos de aplicação de técnicas baseadas em classificadores, que é uma subclasse de métodos algorítmicos apresentados na seção 3.1.2. Estes métodos permitem o uso de múltiplas medidas de similaridade simultaneamente, mas também tem suas desvantagens, discutidas no capítulo 3.

4.2.1 Hasan, Salem e Zaki (2006)

Em Hasan, Salem e Zaki (2006), estudou-se o problema de predição de arestas em redes de co-autoria com o uso de *machine learning*. Os autores descrevem que seu trabalho está estruturado em quatro partes: a) a explicação de como preparar uma base de dados para tratar do problema; b) a identificação de uma lista de características que são tanto importantes para a predição, quanto são pouco custosas em sua obtenção; c) experimentos com conjunto de algoritmos de aprendizado de máquina e respectivos comparativos; d) avaliação de características através de algoritmos de ranqueamento. As duas bases de dados bibliográficas BIOBASE e DBLP, que possuem publicações acerca das áreas de Biologia e Ciência da Computação, foram usadas para gerar grafos de co-autoria. No trabalho buscou-se prever co-autorias para os períodos de 1998 a 2002 e 1990 a 2004 respectivamente.

Para a divisão de treino e teste, separando a base em partes não sobrepostas, escolheu-se pares de autores que publicaram na parte de treino, mas não em conjunto. Assim, cada par destes pode representar um exemplo positivo ou negativo na previsão, dependendo se eles publicaram ou não na parte de testes. Isto torna o problema em classificação binária. As características selecionadas foram divididas em medidas de proximidade, agregadas ou não: quantidade de palavras-chave iguais; soma dos artigos publicados; soma de vizinhos; soma de palavras-chave total; soma do código da classificação; e

soma do *log*(vizinhos secundários). Já as medidas topológicas utilizadas foram: caminho mínimo; índice de clusterização e caminho mínimo em um grafo de palavras-chave. No caso do caminho mínimo, os autores estavam conectados a palavras chave de seus artigos publicados. Os autores também afirmam que utilizaram coeficiente de Jaccard, índice de Adamic/Adar e outras heurísticas comuns, mas não obtiveram resultados relevantes.

Os algoritmos de classificação usados foram: Árvore de decisão; SVM com dois tipos de *kernel*, linear e RBF ; KNN (K-Nearest Neighbours); *Multilayer Perceptron*; Naive Bayes e *Bagging* (*Agregação por Bootstrap*). A plataforma usada para a aplicação dos algoritmos foi o projeto *WEKA*, e para apenas um dos classificadores foi utilizada a linguagem de programação e ambiente *Matlab*, o algoritmo KNN. Dessa forma, os algoritmos testados todos obtiveram acurácia acima de 80%, sendo que SVM com kernel RBF foi a melhor geral para as duas bases. O ranqueamento de atributos chegou a conclusão de que palavra-chave em comum foi o atributo mais significativo, seguido de soma de vizinhos em comum, soma de artigos publicados e caminho mínimo.

4.2.2 Bartal, Sasson e Ravil (2009)

Em Bartal, Sasson e Ravid (2009) foi proposto um método de precisão de arestas baseado em Análise de Redes Sociais - ARS e mineração de texto (*Text Data Mining* - TDM). Na questão TDM, são examinados dois métodos: Processamento de Linguagens Naturais (*Natural Language Processing* - NLP) e modelo de espaço vetorial (*Vector Space Model* - VSM). Estes modelos incluíam características dos artigos publicados para computar semelhança entre os autores.

A base de dados utilizada foi a DBLP referente a publicações na área de Ciência da Computação, durante os anos de 1970 até 1984, onde os primeiros quatro anos foram usados para treino e os subsequentes para testes. As características usadas para treinar o modelo incluem: o modelo VSM; número de co-autorias; vizinhos em comum; coeficientes de Jaccard, Adamic/Adar, conexão preferencial; coeficientes de agrupamento com vizinhança 1 e 2; grau e grau normalizado dos vértices; caminho mínimo e medidas de proximidade e entrelaçamento.

Os classificadores empregados foram Redes Neurais e Árvores de Decisão implementados com SPSS Clementine e C 5.0 respectivamente. Através da análise dos resultados pôde-se observar que a predição com técnicas de SNA podem se beneficiar com

o uso de técnicas de mineração de texto, sendo que os modelos com essa característica obtiveram cerca de 15% a 20% mais acurácia do que os que não a incluíam.

4.2.3 Soares e Prudêncio (2012)

Para prever arestas em duas redes de co-autoria Soares e Prudêncio (2012), calcularam os coeficientes de similaridades entre autores em intervalos diferentes de tempo. Os autores avaliaram a influência dessa similaridade ao longo do tempo tanto em abordagens de classificação supervisionadas quanto em não-supervisionadas. As bases de dados foram retiradas de seções distintas de artigos publicados na área de física da biblioteca digital *Ar-Xiv*.

As características usadas para treinar o modelo são topológicas e contém coeficientes bastante utilizados: Adamic/Adar, Jaccard, Conexão Preferencial e Vizinhos em Comum. Foram empregados múltiplos modelos de previsão não-supervisionada: Regressão Linear, Caminhada Aleatória (*Random Walk*), Suavização Exponencial Simples, Suavização Exponencial Linear, e Média de Movimentação nos intervalos. Apenas um modelo de aprendizado supervisionado foi aplicado, SVM através da plataforma *Weka*. Com isso, observou-se que modelos podem se beneficiar de características temporais tanto no aspecto supervisionado quanto no não-supervisionado. Da mesma forma, constatou-se que essas características reduzem a necessidade de um histórico temporal forte, beneficiando modelos que utilizam valores mais recentes para as previsões.

4.2.4 Yu et al. (2014)

Yu et al. (2014) aplicaram técnicas de predição de arestas topológicas gerais e as compararam com técnicas topológicas individuais, em redes de co-autoria da área biomédica. A base de dados utilizada foi a da *Web of Science* (WoS), com artigos na área especifica de doenças de artérias coronárias. As características topológicas usadas foram: Vizinhos em Comum; coeficientes de Jaccard, Adamic/Adar e conexão preferencial; as propriedades baseadas em caminho foram Katz e Propflow.

Os classificadores usados foram SVM e Regressão Linear, implementados na plataforma Weka. O método *Wrapper*, presente na plataforma foi utilizado para a seleção das melhores características para previsão, no qual fazem-se buscas entre subconjuntos

de características. Para análise dos resultados, os classificadores foram comparados a predições aleatórias, para os quais ambos tiveram um aumento de performance de cerca de 20%. Foram também avaliados autores com base em sua prolificidade, separando-os em 5 grupos por número de artigos publicados, e observou-se que os modelos obtiveram boa precisão na maioria dos grupos com exceção a autores que publicaram pouco. O método de seleção de características obteve que Adamic/Adar, Conexão Preferencial, Katz e Propflow tiveram as melhores pontuações para Regressão Linear, e para SVM, as mesmas características foram encontradas substituindo-se apenas Katz com Vizinhos em Comum.

4.2.5 Julian e Lu (2015)

Em seu trabalho Julian e Lu (2015) aplicam algoritmos de aprendizado de máquina supervisionado usando características topológicas e heurísticas do grafo. Foram utilizadas uma série de bases de dados, que incluiam cinco redes de co-autoria do *arXiv*, que é um repositório eletrônico para artigos para principalmente a área de Ciências Exatas. Essas redes estavam publicamente disponíveis através do Projeto de Análise de Redes de Stanford (*Stanford Network Analysis Project - SNAP*). Para cada uma das redes analisada foi aplicado um pré-processamento, que consistia em: remover os nodos com grau menor ou igual a 3, visto que esses possuiam menos probabilidade de auxiliar na previsão; selecionar e remover 10% das arestas para simular um ponto de tempo diferente da rede e poder avaliar o crescimento da mesma; aplicar divisão da base em treinamento e teste na proporção 70-30.

O objetivo era a partir da rede com arestas removidas G_0 prever as arestas na rede atual G_n. Das características usadas para a aplicação do algoritmo, oito referiam-se a relacionamento do grau entre dois vértices, mais oito referiam-se a participação de grupos parciais, frequência em grupos de 1,2,3 e 4 membros, e o relacionamento entre a frequência e o número de vizinhos em comum. Também foram consideradas quatro características heurísticas, vizinhos em comum, coeficiente de Adamic/Adar, coeficiente de Jaccard, e o coeficiente de conexão preferencial.

Os algoritmos de aprendizado de máquina foram desenvolvidos com a biblioteca Python Scikit-Learn, sendo eles três: regressão logística, florestas aleatórias, e rede neural. Com isso, para avaliar os resultados dividiram-se as características utilizadas

em subconjuntos de cada tipo e o conjunto total. Observou-se que os três classificadores obtêm acurácia próxima de 80% para o conjunto total, e que as medidas heurísticas isoladamente alcançavam cerca de 75% de acurácia do conjunto total, sendo que as características de proximidade e grau chegaram muito próximas do conjunto total. Através do uso de características topológicas, e não de domínio, pôde-se observar similaridades na avaliação de performance entre as redes.

4.2.6 Maruyama e Digiampetri (2016)

Para prever co-autorias Maruyama e Digiampietri (2016) utilizaram informações da plataforma Lattes de 637 pesquisadores que foram professores em programas de graduação em Ciência da Computação durante o período de 2004 à 2009. Os autores abordaram dois aspectos na previsão de redes de co-autoria: o problema tradicional de predição de arestas, ou seja, a predição de arestas independente da sua existência no presente, e a predição de apenas arestas que não existem no presente, ou seja a diferença é sutil, mas o novo problema consiste em desconsiderar co-autorias que existem no presente para a avaliação. O intervalo de tempo usado para as previsões foi de 1971 a 2015, sendo que os intervalos de 1971-2000, 2001-2005, 2006-2010 foram considerados como passado, presente e futuro respectivamente para o treinamento dos modelos, e para testes os intervalos 1976-2000, 2006-2010, 2011-2015.

No trabalho, 31 características foram extraídas ou calculadas para a aplicação em múltiplos classificadores da plataforma Weka, sendo que dessas 16 eram atributos dos pares ou dos pesquisadores (retirados da plataforma Lattes) e 15 eram características estruturais do grafo. Um filtro foi aplicado para autores que possuíam potencial muito baixo de virar co-autores, o que resultou na remoção de em torno 200.000 pares. Utilizou-se da plataforma *Weka* para aplicar uma variedade de classificadores, dos quais para o problema tradicional de predição de arestas o *Attribute Selected Classifier* se destacou, e para o problema de predição de novas arestas o classificador *BayesianLogisticRegression*. No entanto, os autores concluem que um classificador que prevê que não há novas conexões tende a ter uma diferença não muito grande destes classificadores para o problema em questão, e que o problema de predição de novas arestas não produziu resultados satisfatórios.

4.3 Considerações sobre o Capítulo

A Tabela 4.1 relaciona os trabalhos citados nas seções anteriores para os quesitos medidas de similaridade, classificadores, ferramentas usadas e tipo de avaliação. Através da análise de nove trabalhos que realizam experimentos de predição de arestas em redes de co-autoria, pode-se observar a variabilidade de uma série de aspectos. Com exceção de algumas como CN, JC e AA, o uso das medidas de similaridade não são consistentes, no sentido de não aparecerem em todos os trabalhos. Da mesma maneira, o uso de classificadores segue o mesmo caminho, com autores avaliando de um a mais de dez classificadores diferentes em seus trabalhos. No quesito ferramentas, observou-se a dominância da plataforma Weka, provavelmente associada à disponibilidade e facilidade de uso da mesma. Outro fator a debater são as métricas aplicadas, onde também não há consenso, com alguns autores limitando seu trabalho a apenas uma métrica e outros sendo mais abrangentes.

A análise destes trabalhos permitiu identificar medidas de similaridade eficazes, bem como complicadores do uso de classificadores. Um destes complicadores ainda pouco tratados na literatura é a questão do desbalanceamento de classes, presente no problema de predição de arestas para redes de co-autoria. Além disto, os trabalhos analisados deixam algumas outras lacunas. Primeiramente, a questão da seleção de variáveis, que neste contexto são as medidas de similaridade apresentadas no capítulo anterior. Autores dos trabalhos relacionados muitas vezes dão enfoque na performance do classificador e não levam em consideração a fase do pré-processamento. Esta consiste no processo de selecionar um subconjunto de características usadas para aplicação no algoritmo de aprendizado de máquina. Estratégias que tratam desta questão tem benefícios como melhores performances de previsão, redução de dimensionalidade, que levam a menores tempos de treinamento, bem como interpretabilidade e generalização do modelo (JAMES et al., 2014). Outra questão pouco avaliada, é a questão da recência das colaborações, isto é, saber quão recente dois pesquisadores mantiveram colaborações. Acredita-se que esta informação seja importante, pois sabe-se que alguns pesquisadores param de publicar com o passar do tempo (MENDONÇA, 2017). Assim, o algoritmo quando alimentado por tais informações poderá identificar se determinadas colaborações são de fato úteis para a previsão. O trabalho a ser proposto no Capítulo 5 visa preencher estas lacunas.

Tabela 4.1: Lista de Trabalhos Relacionados

Trabalho	**Medidas de Similaridade**	**Classificador**	**Plataforma**	**Avaliação**
Liben-Nowell e Kleinberg (2003)	CN, PA, JC, AA, SimRank, Hitting time, PageRank, Katz	Não	-	Precisão
Brandão et al. (2013)	CORALS, Affin, GLI	Não	-	Acurácia, Diversidade, Novidade, Cobertura
Gao et al. (2014)	JC, PA, AA, CS, HPI, HDI, LHN, Katz, Sorensen	Não	-	AUC
Hasan, Salem e Zaki (2006)	Atributos de publicação, CM, JC, AA	SVM, DT, KNN, RN, NB, Bagging	WEKA, Matlab	Acurácia, Precisão, Revocação, Medida F
Bartal, Sasson e Ravid (2009)	TDM, VSM, JC, AA, PA, CM, Medidas de ARS	RN, DT	SPSS Clementine, C 5.0	Acurácia
Soares e Prudêncio (2012)	AA, JC, PA, CN	SVM	WEKA	AUC
Yu et al. (2014)	CN, JC, AA, PA, Katz, Propflow	Regressão Linear, SVM	WEKA	Acurácia, Precisão, Revocação, Medida F, AUC
Julian e Lu (2015)	Tríades, CN, AA, JC, PA	Regressão Logística, RF, RN	Python scikit-learn	Acurácia
Maruyama e Digiampietri (2016)	Informações do Lattes, JC, AA, RA, HPI, HDI, Katz, CM, Salton, Sorensen e outras	Muitos classificadores	WEKA	Precisão, Revocação, Medida F, AUC

5 Proposta

Este capítulo apresenta a proposta inicial deste trabalho, que foi definida com base na fundamentação teórica e nos trabalhos relacionados estudados. O trabalho proposto busca abordar o problema de previsão de arestas para uma rede de co-autoria composta por pesquisadores que publicaram no Simpósio Brasileiro Sobre Fatores Humanos em Sistemas Computacionais (IHC) através do uso de conjuntos de técnicas de medida de similaridade e técnicas baseadas em classificadores. O objetivo principal é avaliar e comparar os fatores que levam a novas colaborações, assim como mensurar o desempenho das diversas técnicas aplicadas. Como resultado, espera-se que seja possível obter uma melhor compreensão dos fatores que levaram à evolução deste grupo de colaboradores.

5.1 Visão Geral

Conforme visto, existem diversos fatores que podem influenciar na evolução de redes de colaboração científica, e o estudo destes nos permite entender comportamentos que podem ou não levar a co-autorias em determinada área. O objetivo do problema de predição de arestas é entender os aspectos relacionados à formação de novas colaborações, usando uma combinação de características estruturais da rede e características específicas dos autores e publicações.

Alguns fatores diferem este trabalho dos demais na área. Primeiramente o foco é a seleção do subconjunto de características de similaridade que melhor representa os dados, não só a avaliação de diversos classificadores para o mesmo conjunto de entradas, como visto na maioria dos trabalhos do capítulo anterior. Associado a isso, este trabalho busca incluir e avaliar características temporais na predição para considerar a recência das colaborações. Também serão avaliadas abordagens para o tratamento da questão do desbalanceamento de classe, pouco abordada nos trabalhos relacionados. Busca-se ainda evitar uma avaliação incompleta, que tornaria o trabalho tendencioso> Isso pode ser evitado através do uso todas as medidas apresentadas nos capítulos 3 e 4, unido a uma análise de cada uma destas, que possibilitará um maior entendimento na viabilidade

de tal medida.

Outro fator é a expansão dos trabalhos publicados por Gasparini e colegas (2013, 2014, 2015, 2016, 2017), através da análise dos autores prolíficos usados em (MENDONÇA, 2017), sob o ponto de vista de predição de arestas. Para tanto, tornando visual a evolução dos grafos de co-autoria para a comunidade de IHC no Brasil, e avaliando-a em termos de predição de arestas. Dessa forma, pretende-se fornecer uma melhor compreensão do desenvolvimento da comunidade, bem como contribuir para a área de predição de arestas que abrange muito mais do que apenas redes de co-autoria. O estudo dos complicadores ligados a cada técnica, como alto desbalanceamento de classe para os métodos baseados em classificadores também trará conhecimentos sobre as melhores estratégias para lidar com o mesmo nesta instância do problema.

5.2 Escolha de Métodos de Predição

Como visto, pretende-se combinar técnicas de predição de arestas baseadas em medidas de similaridade e baseadas em classificadores. Dada a variedade de medidas de similaridade que podem ser aplicadas às técnicas, a escolha de quais medidas utilizar envolverá um um trabalho de criação e extração de características, bem como uma fase de pré-processamento para a seleção das mais relevantes.

Quanto às técnicas baseadas em similaridade, as abordagens podem ser divididas em baseadas em vizinhança, caminho, caminhada aleatória e por atributos. Para cada subclasse destas 4 opções, será escolhida ao menos uma abordagem para cada. O número de abordagens a avaliar de cada subclasse será dependente de alguns fatores principais: suporte nativo em ferramenta, relevância da medida baseada em trabalhos relacionados, e relevância baseada em métodos de seleção de variáveis. Medidas com alta probabilidade de serem avaliadas são as medidas heurísticas mais utilizadas, como índice de Adamic/Adar, coeficiente de Jaccard, vizinhos em comum e Katz. Nota-se que estão inclusos também os métodos baseados em atributos, que normalmente não são usados isoladamente, mas serão úteis para a aplicação de técnicas baseadas em classificadores. Com relação a estes dados, pode-se extrair várias informações de pares interessantes, como por exemplo: palavras-chave em comum em publicações, número de publicações conjuntas, afiliação, localização e outros.

Sobre as técnicas baseadas em classificadores, primeiramente serão selecionados os algoritmos a serem aplicados para os quais as opções são vastas, dado o amplo suporte disposto nas ferramentas comumente usadas como *scikit-learn* e a plataforma Weka. Os critérios para isso se darão pela relevância e performance nos trabalhos relacionados, bem como suporte em ferramentas. Outro fator de decisão são os conjuntos de características aplicadas a cada método, para as quais a ênfase será dada aos métodos que obtiveram melhor performance nas medidas de similaridade e por atributos.

5.3 Metodologia de Avaliação dos Resultados

Busca-se encontrar a melhor combinação de características que obtêm os melhores resultados de acordo com as métricas de avaliação apresentadas anteriormente, e através da realização de experimentos e comparativos encontrar-se-a os fatores que estão mais correlacionados a colaborações no contexto do evento IHC.

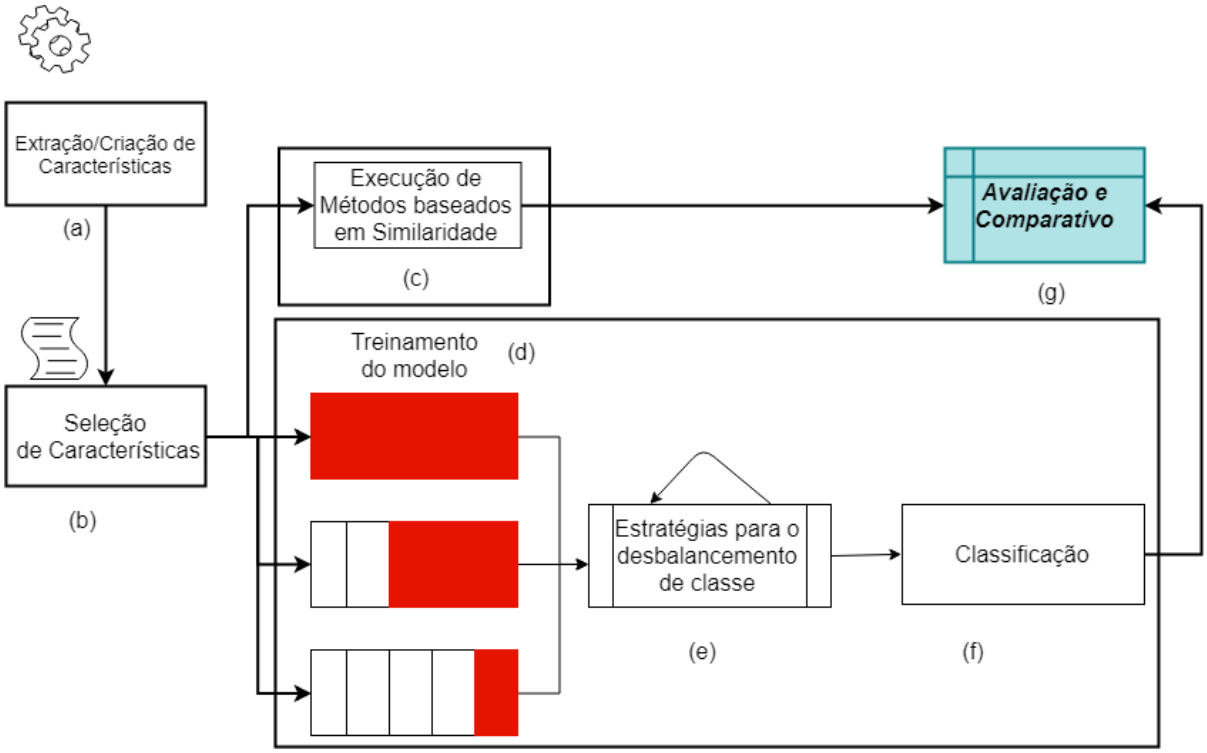
Mais especificamente, em dado intervalo para o qual existem dados de colaborações para o evento, serão definidas faixas de tempo dentro deste intervalo para treinamento e testes, as quais servirão de dados de entrada para as técnicas utilizadas. Dos dois tipos de técnica, o primeiro, baseado em apenas medidas de similaridade individualmente, será utilizado como base para a escolha e avaliação dos conjuntos de medidas para o segundo, baseado em aprendizado de máquina. Para que assim, estas sejam aplicadas, comparadas e devidamente avaliadas, para a extração de informações relevantes à formação de novas co-autorias ou fortalecimento das pré-existentes.

Um número de experimentos será considerado como: serão geradas e comparadas situações com o uso de técnicas para o balanceamento de classes e sem uso dessas técnicas, com o uso de atributos externos como dados de palavras chave e sem o uso destes, com o uso de características associadas ao tempo quanto não associadas ao tempo para avaliar a questão da recência das colaborações. Outro teste para a avaliação da recência, através do uso de testes com diferentes intervalos de treinamento, para verificar a influência disso no modelo.

Para computar a assertividade dos métodos de previsão, serão aplicadas todas as métricas de avaliação apresentadas no Capítulo 2. Esta prática será adotada para evitar uma avaliação tendenciosa, e para que seja possível selecionar e recomendar os modelos e

respectivos métodos que obtiveram a maior assertividade.

Figura 5.1: Estrutura de resolução do problema proposta



Fonte: O autor

A figura 5.1, mostra um esquema de como a proposta deste trabalho será aplicada. O fluxo inicia-se em (a) com a criação e extração de características, estas serão as medidas de similaridade que possivelmente serão aplicadas ao modelo, o que inclui medidas baseadas em topologia como vizinhança, caminho e caminhada aleatória, bem como medidas externas a topologia ou de atributos como localização geográfica e palavras-chave em comum. Na fase (b), ocorrerá a seleção de características, onde um subconjunto das medidas extraídas ou geradas será selecionado através de técnicas de pré-processamento para a redução de dimensionalidade, que encontrarão as características potencialmente mais importantes para o modelo.

Para a fase (c), serão computados um método representante de cada subclasse de métodos baseados em similaridade para que estes sejam utilizados como base para o comparativo com métodos baseados em classificadores. As fases (d), (e) e (f) estão associadas aos métodos baseados em classificadores. Em (d) serão gerados os conjuntos de treinamento, ou seja as medidas de similaridade para os anos anteriores a previsão. Esta fase lida com a recência das colaborações, visto que serão gerados conjuntos de treinamento reduzindo iterativamente o número de anos anteriores presentes na base, ou seja,

treinamento com os anos: 1998 a 2014, 2002 a 2014, 2006 a 2014, 2010 a 2014 e 2012 a 2014. Para cada dos conjuntos gerados em (d), são aplicadas técnicas para o balanceamento das classes em (e), que são técnicas de amostragem, como reduzir número de dados de treinamento para os casos de maioria até que haja equilíbrio com os casos de minoria (*undersampling*), ou clonar casos de minoria até que haja equilíbrio (*oversampling*), ou uma combinação de ambos usando de uma técnica baseada em vizinhos em comum, *Synthetic Minority Over-sampling Technique - SMOTE*, (BOWYER et al., 2011), e ainda os testes com a base não balanceada. Tais técnicas serão melhor explicadas no TCC2. Para parte (f), serão escolhidos apenas dois classificadores, SVM e Redes Neurais, pois estes são os que mais frequentemente aparecem em trabalhos relacionados, como visto no capítulo 4.

Por fim em (g), haverá comparativos entre as técnicas, no contexto de classe da técnica com métodos baseados em similaridade contra métodos baseados em classificadores, no contexto recência dos dados e no contexto balanceamento de classe. Tais comparativos incluirão todas as métricas de avaliação apresentadas anteriormente, e permitirão encontrar as estratégias e medidas de similaridade que obtêm melhor performance para a base de dados avaliada. Dessa forma, encontrando os fatores associados a formação de conexões e a evolução da rede ao longo dos anos.

5.4 Base de Dados do IHC

A área de Interação Humano-Computador (IHC) é um campo de pesquisa que estuda como as pessoas interagem com os sistemas computacionais e até que ponto os computadores são ou não desenvolvidos para uma interação bem sucedida com os seres humanos (JONES, 2016). A área tem como um de seus objetivos investigar e produzir alternativas tecnológicas envolvidas no design e na avaliação de interfaces de usuários, para as pessoas interagirem de forma produtiva com métodos, técnicas, modelos e representações utilizados em diversos sistemas (SOUZA, 2006).

Trabalhos como os de Gasparini, Kimura e Pimenta (2013) realizaram abordagens bibliométricas, como número de artigos publicados por ano, principais autores, instituições e até redes de co-autoria para análise da comunidade de IHC brasileira. Porém, nenhuma análise preditiva ou de recomendação foi estudada neste contexto.

Alguns dos autores de Mendonça (2017) mantêm um histórico de publicações do evento IHC desde 1998 em um banco de dados relacional no *MySQL*. Este banco contém dados sobre as publicações, como autores, resumo, palavras-chave, assim como dados específicos de autores como nome, gênero, artigos publicados, e ainda dados complementares extraídos da plataforma *Lattes*. No entanto, a base necessitava de atualizações visto que a última ocorreu em 2015.A base também não inclui artigos resumidos, de menos de 4 páginas, até o momento. Entretanto, um mecanismo de *scraping* de dados sobre os anais do IHC disponíveis na biblioteca digital da ACM foi construído pela Professora Simone Barbosa. Os dados coletados correspondem aos artigos completos e resumidos para os anos de 2006-2017. Este trabalho foi encarregado da tarefa de atualização da base de dados, e atualmente a situação encontra-se em status parcial, com a atualização apenas dos artigos completos até 2017. Em ato contínuo ainda devem-se incluir os artigos curtos para todos os anos.

5.5 Resultados Iniciais

Durante o período inicial deste projeto, parte do tempo foi dedicado a atualização da base MySQL até 2017, portanto algumas informações sobre a base já puderam ser extraídas e algumas visualizações puderam ser geradas. Estas serão destacadas a seguir, com o intuito de visualizar a evolução das redes de colaboração do IHC ao longo dos anos.

Após atualização dos dados básicos de co-autoria para o banco de dados, foi possível gerar grafos de co-autoria via *Python* e a biblioteca *networkX* bem como as respectivas visualizações pela ferramenta *Gephi*. As Figuras 5.2, 5.3, 5.5, 5.7 e 5.9 representam respectivamente as redes de co-autoria dos anos 1998, 2002, 2010, 2014 e 2017. Vale lembrar que cada nodo constitui um autor de artigo completo, e cada aresta é uma representação de uma co-autoria entre dois autores. Para melhor visualização, nodos com maior grau estão destacados por cores mais fortes e tamanhos maiores. As Figuras 5.2, 5.4, 5.6, 5.8 e 5.10, demonstram a distribuição de grau para cada ano mostrado, que é a contagem de nodos por grau no grafo avaliado.

Figura 5.2: Grafo de co-autorias do IHC em 1998 (esquerda), e Distribuição de grau (direita)

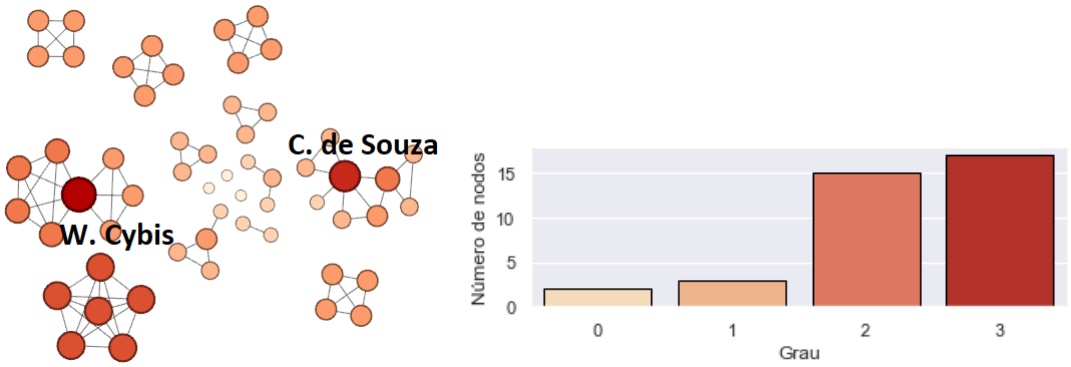
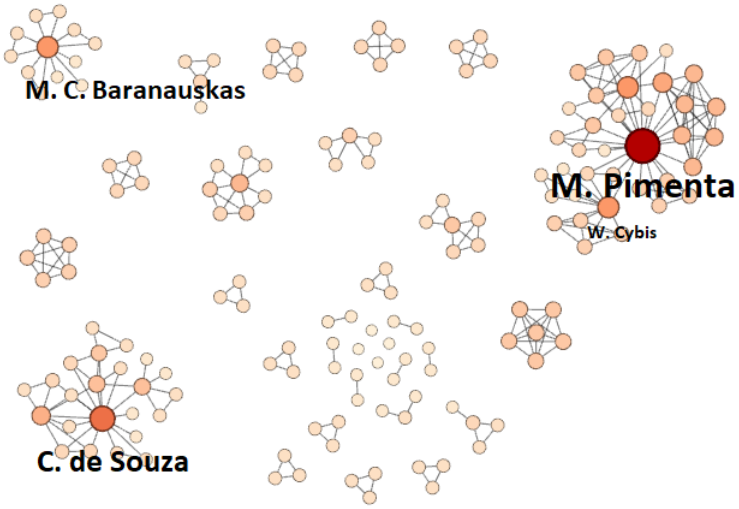
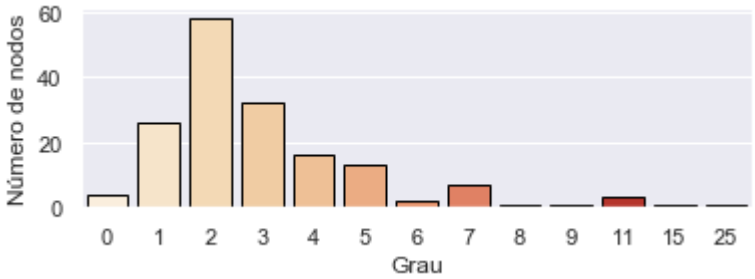


Figura 5.3: Grafo de co-autorias do IHC em 2002



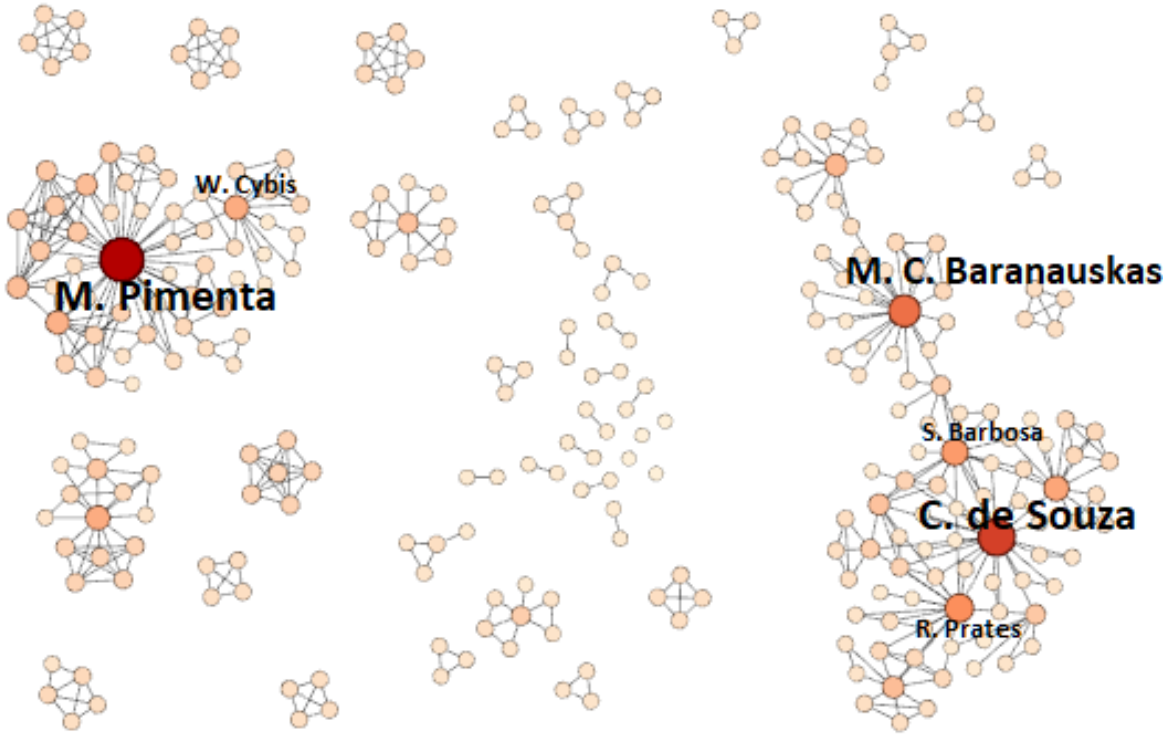
Fonte: O autor

Figura 5.4: Distribuição de grau em 2002



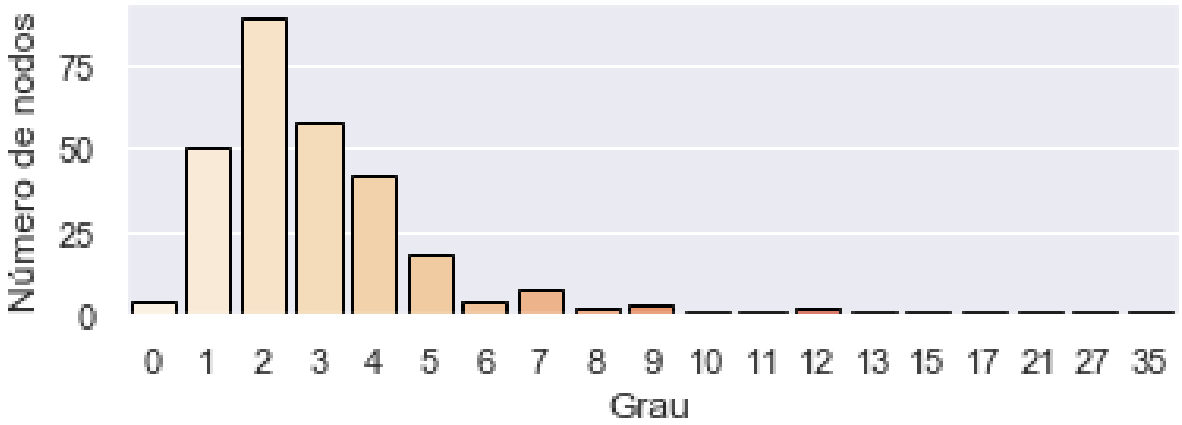
Fonte: O autor

Figura 5.5: Grafo de co-autorias do IHC em 2010



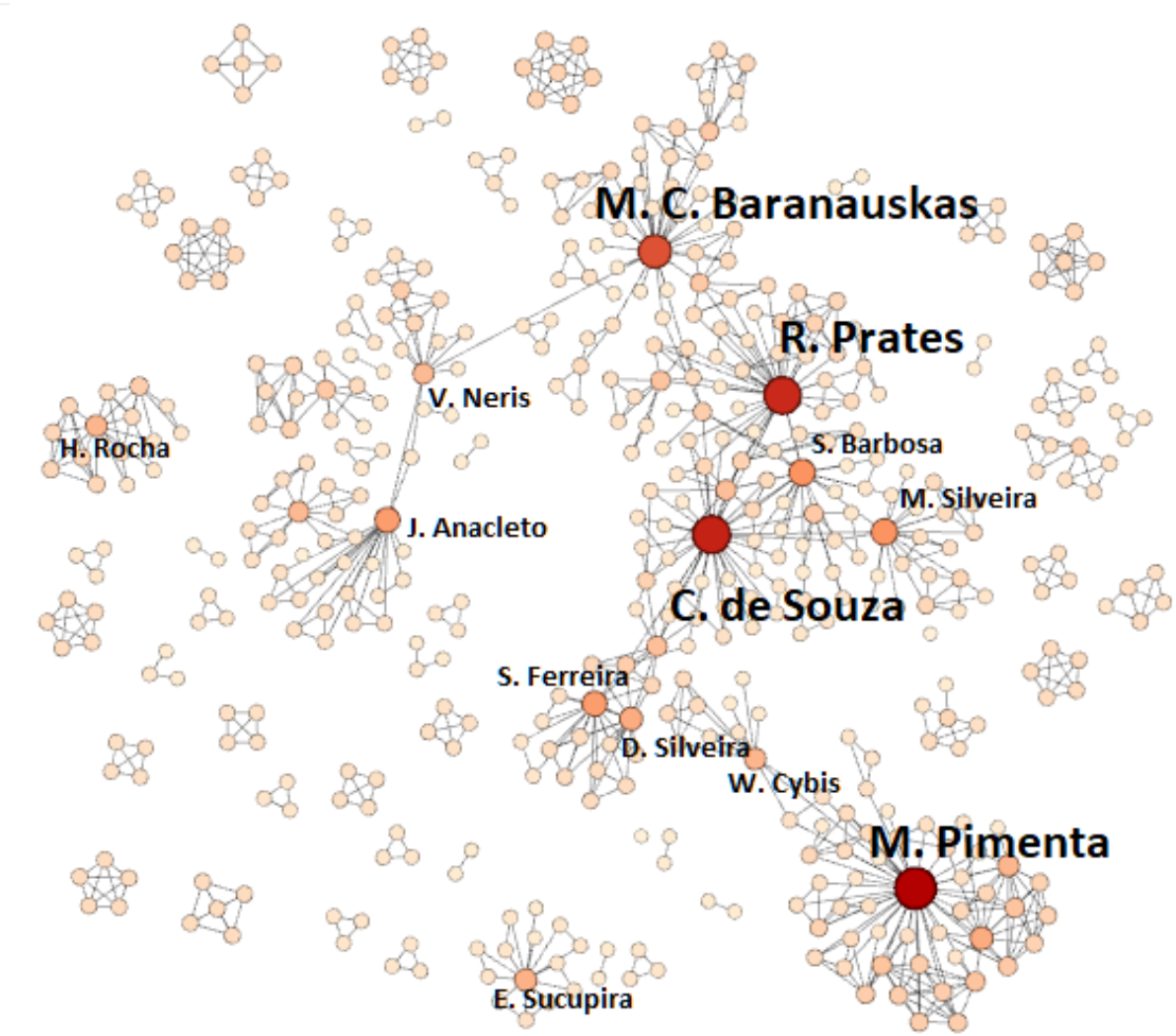
Fonte: O autor

Figura 5.6: Distribuição de grau em 2010



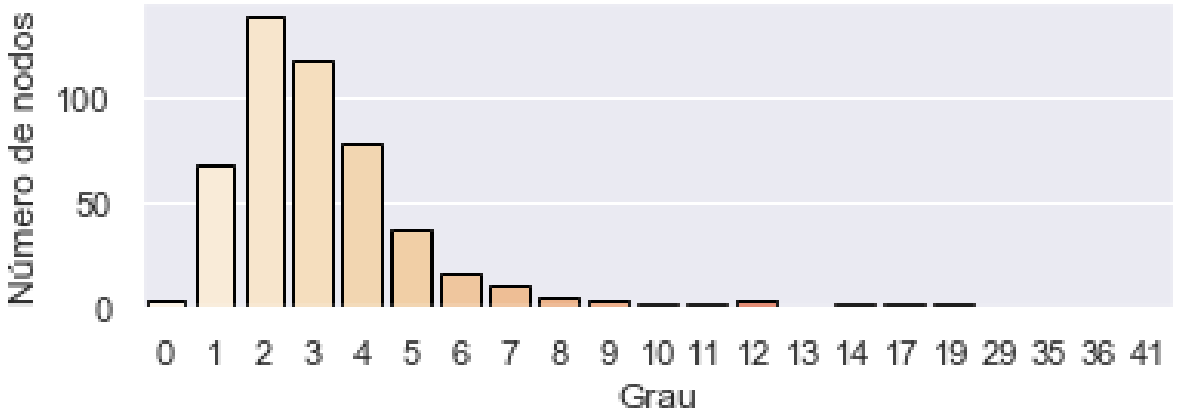
Fonte: O autor

Figura 5.7: Grafo de co-autorias do IHC em 2014



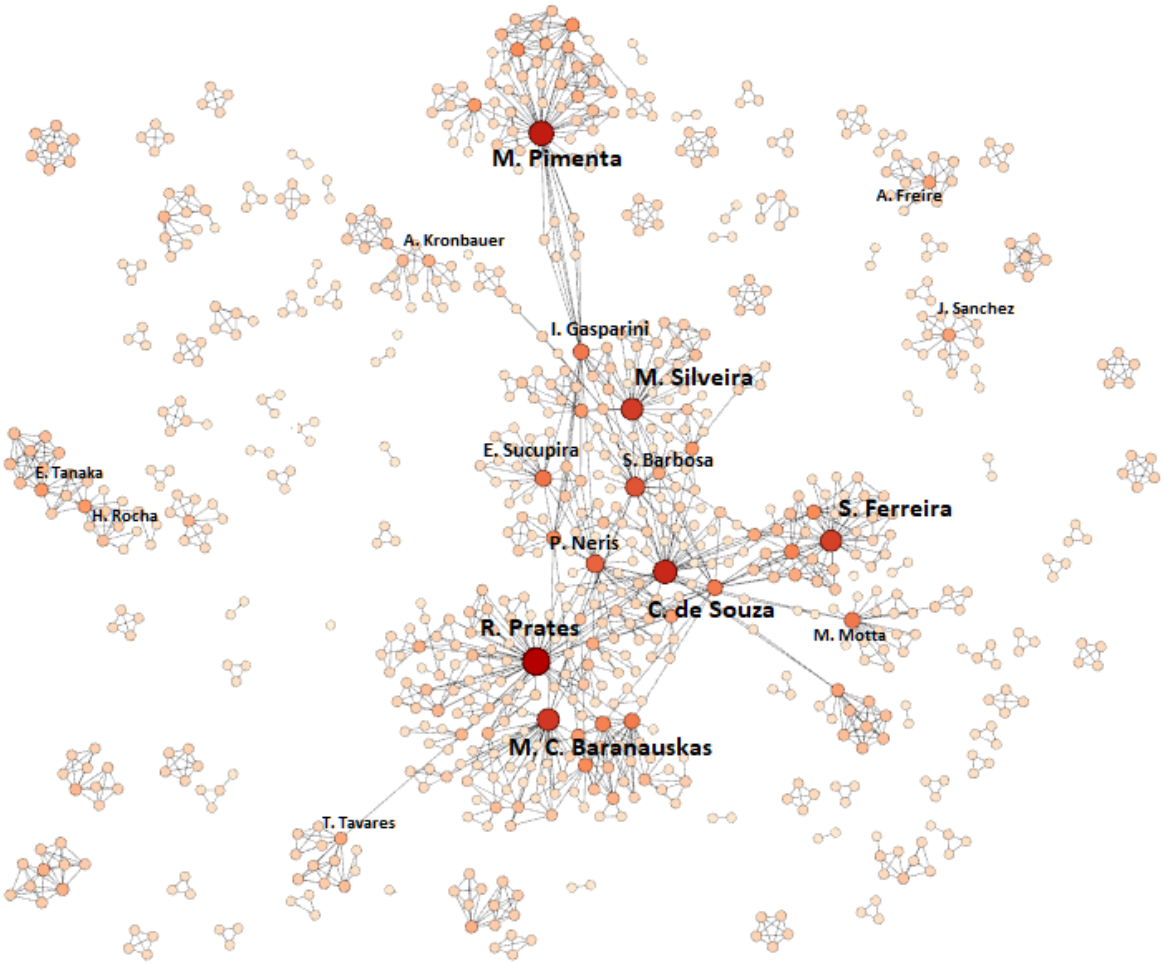
Fonte: O autor

Figura 5.8: Distribuição de grau em 2014



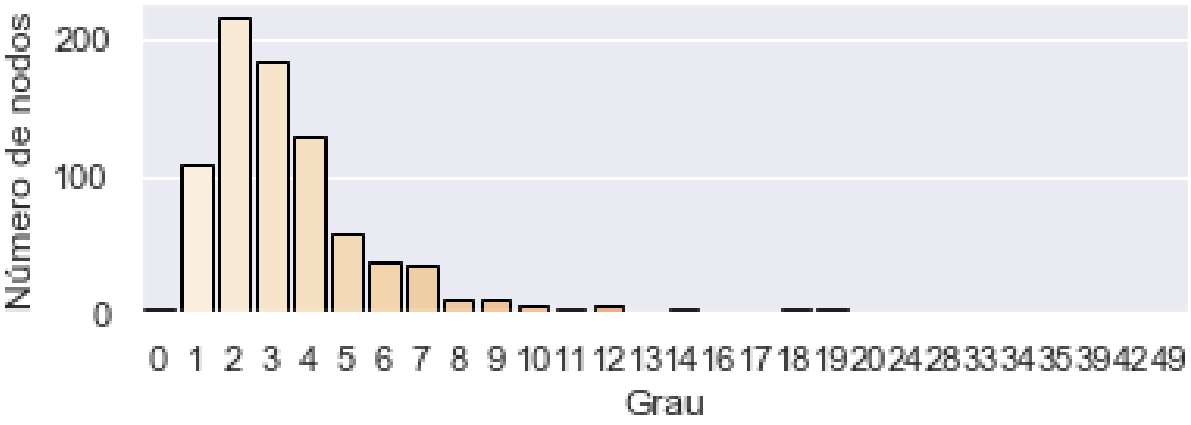
Fonte: O autor

Figura 5.9: Grafo de co-autorias do IHC em 2017



Fonte: O autor

Figura 5.10: Distribuição de grau em 2017



Fonte: O autor

Nota-se a evolução da comunidade ao longo dos anos, como visto, o primeiro ano do evento mostrou uma distribuição de grau relativamente uniforme diferente das

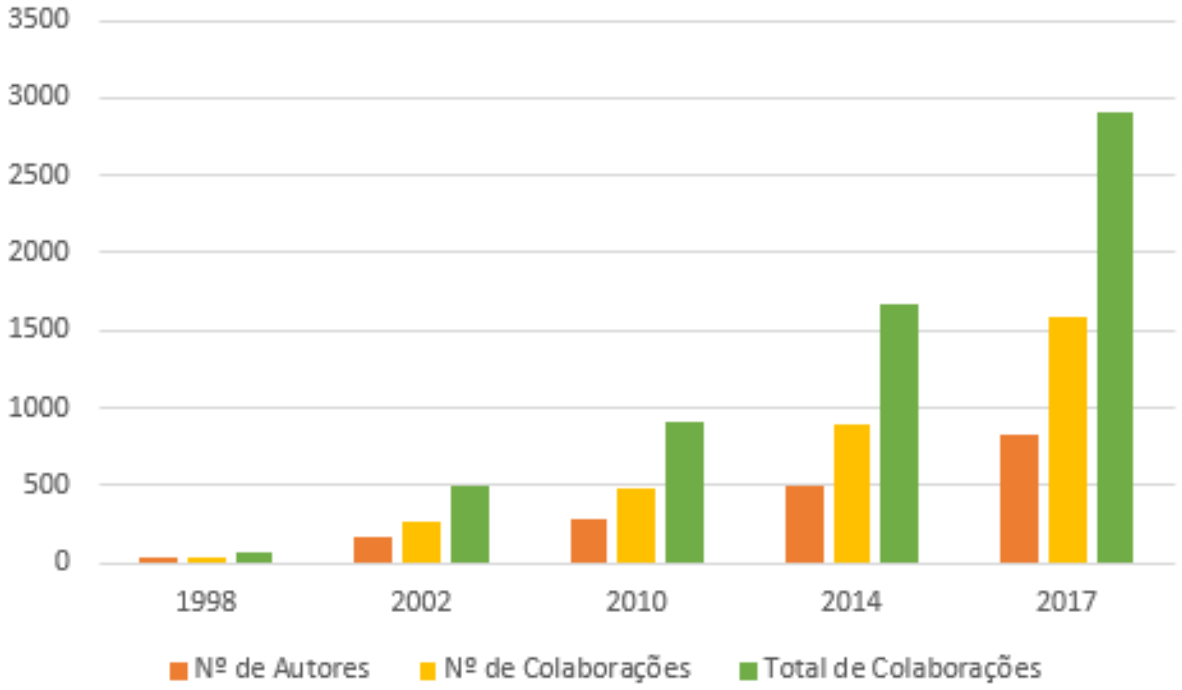
características de redes complexas. Porém quatro anos após, em 2002, já é possível notar a característica de redes livres de escala, com um número pequeno de nodos com maior influência na rede. Durante esses anos iniciais alguns autores se destacam; Em 1998 destacaram-se os autores Walter A. Cybis e Clarisse S. de Souza. Já em 2002 surgem os pesquisadores Maria Cecília C. Baranauskas e Marcelo S. Pimenta, enquanto a professora Clarisse de Souza mantém sua prolificidade. A partir deste ano a distribuição de grau começa a tomar um caráter de livre de escala, com poucos nós uma quantidade grande de conexões e muitos com uma quantidade reduzida.

Em 2010, as estruturas das professoras Maria Cecília Baranauskas e Clarisse de Souza se conectam. Além disto, é possível notar autores menores dentro dessa nova estrutura começando a se destacar, como Simone Barbosa e Raquel Prates. Alguns anos depois, a comunidade já está bastante difusa e complexa, pois inúmeros autores se destacam. Um fato interessante é que o grupo maior, dominada pelos autores que se destacaram em 2010, ainda não está conectada ao grupo do professor Marcelo Pimenta. Entretanto, esta situação muda em 2017, com professora Isabela Gasparini sendo a conexão que une o grande grupo ao do grupo do Professor Marcelo Pimenta. A partir destes grafos é possível observar inúmeros autores em destaque, tanto em grupos desconexos quanto pertencentes a grupos maiores e mais conectados.

Algumas observações podem ser feitas, como por exemplo, um grafo que simula um sistema real, não necessariamente é livre de escala desde seu início, pelo menos não neste caso, mas a tendência é adquirir a característica ao longo do tempo. Outra observação é que há uma tendência da formação de grupos isolados, que eventualmente, ao longo de seu crescimento, acabam por se conectar a grupos maiores e mais conectados. Além disto, nota-se que alguns grupos realizam publicações por um período e depois param, sendo essa uma das razões pelo grande número de grafos pequenos e isolados.

Essas foram algumas observação obtidas de uma análise preliminar visual da rede, alguns outros dados complementares estão dispostos na Figura 5.11.

Figura 5.11: Evolução da rede ao longo dos anos, no quesito colaboração



Fonte: O autor

Considere que nos grafos de co-autoria, uma aresta é estabelecida entre dois pesquisadores (nodos do grafo) sempre que há pelo menos uma publicação em conjunto entre eles. Além disto, as arestas são ponderadas com o número de publicações em que ambos colaboraram. Desta forma, o gráfico da Figura 5.11 mostra em laranja, o número de pesquisadores que publicou no evento para o devido ano, isto é, a quantidade de nós do grafo. As barras em amarelo mostram o número de arestas do grafo de co-autoria, ou seja o número de pares de autores diferentes que publicaram em conjunto. Nas barras verdes, o número total de colaborações para o ano é representado, incluindo reincidências. As reincidências correspondem à soma das ponderações de todas as arestas do grafo de co-autoria.

A partir deste gráfico é possível observar um crescimento quase que exponencial na evolução da rede de co-autoria do IHC. Não só isso, mas no quesito colaboração, observamos que o número de colaborações reincidentes é quase tão grande quanto o número de colaborações únicas, o que mostra que reincidências podem ser fatores interessantes a se considerar. Outro padrão observado, é o número de co-autorias em relação ao número de autores, que no início está equilibrado, mas com o passar do tempo, o número de co-autorias aumenta a um passo mais rápido, e em 2017 torna-se quase o dobro do que o

número de autores.

Algumas medidas ainda não foram computadas, mas serão aplicadas para a formalização de detalhes desta comunidade, como o coeficiente *small* que mede o quando uma rede é de pequeno mundo e a medida de escala da rede. Outra medida de interesse é o número de co-autorias possíveis ao longo do tempo em proporção as concretizadas.

5.6 Considerações sobre o capítulo

Este capítulo forneceu uma visão geral do que foi observado ao longo dos capítulos anteriores em termos de predição de arestas. Neste contexto, foram identificadas algumas características das redes de co-autorias do IHC ao longo dos anos. Assim como, a proposta de como este trabalho irá lidar com a lacunas apontadas no capítulo 4. Alguns resultados também iniciais foram apresentados, principalmente ligados à atualização das bases de dados e geração de visualizações para a rede. Por fim, este capítulo forneceu algumas análises sobre estas redes no quesito de distribuição de grau dos nós, e número de colaborações ao longo do tempo.

6 Conclusões Parciais

Este trabalho, inicialmente, buscou realizar um levantamento de técnicas para a abordagem do problema de predição de arestas, por isso fez-se necessária uma contextualização do problema. Os conceitos de redes complexas, suas subclasses, e redes de colaboração científica foram introduzidos, para que então pudessem ser apresentadas as teorias sociais usadas como bases para predição de arestas. Neste contexto, foram observadas conceitos como redes de pequeno mundo e redes sem escala, que por sua vez estão presentes em redes colaborativas de co-autoria, que são o foco específico deste trabalho.

A importância da recomendação ou previsão de novas conexões para redes de co-autorias foram apresentadas, bem como os problemas e técnicas associados a isso. De acordo com o quesito técnicas, foram destacados dois tipos, abordagens que usam apenas medidas de similaridade e abordagens que usam de classificadores, ambas tendo suas vantagens e desvantagens. Junto a isso, foram apresentadas métricas para a avaliação dos modelos de predição e ferramentas comumente utilizadas na área.

Como visto, a vantagem principal do uso de apenas medidas de similaridade está na facilidade de computação e avaliação, porém a desvantagem dessas medidas é que as previsões são realizadas com base na medida isoladamente, sem o auxílio de outras informações. Já classificadores tem como principal vantagem o uso de múltiplas características de similaridade simultaneamente, todavia possuem desvantagens como a questão do desbalanceamento de classe que dificultam e comprometem a avaliação destes modelos.

Nove trabalhos relacionados foram analisados, os quais realizam experimentos de predição de arestas em redes de co-autoria. Através desta análise pode-se observar a variabilidade de uma série de aspectos. Várias medidas de similaridade não são universalmente utilizadas, e a escolha destas varia conforme o trabalho analisado. Da mesma maneira, a escolha de classificadores, com autores avaliando de um a mais de dez classificadores diferentes em seus trabalhos. No quesito ferramentas, observou-se a dominância da plataforma Weka, provavelmente associada à disponibilidade e facilidade de uso da mesma. Outro fator a debater são as métricas aplicadas, onde também não há consenso, com alguns autores limitando seu trabalho a apenas uma métrica e outros sendo mais

abrangentes. Através disso, foi possível observar algumas lacunas abertas. A questão da seleção de variáveis, ou seja as medidas de similaridade aplicadas, a recência das colaborações, e a previsão com base em atributos temporais. Além disto, uma lacuna observada foi o não tratamento do desbalanceamento de classes, além de uma avaliação completa com base nas métricas apresentadas nos trabalhos.

Por fim, foi demonstrado como o trabalho proposto visa preencher com mais propriedade as lacunas identificadas nos trabalhos relacionados. Bem como alguns resultados iniciais, principalmente ligados à atualização das bases de dados e geração de visualizações para a rede. Algumas análises sobre estas no quesito de distribuição de grau ao longo dos anos, e número de colaborações ao longo do tempo. Dessa forma, auxiliando na compreensão do modelo de evolução da comunidade de IHC, assim como contribuindo com a área de predição de arestas que abrange muito mais do que apenas redes de co-autoria.

6.1 Cronograma

Com relação ao cronograma proposto no plano de trabalho, poucas alterações ocorreram, elas aconteceram devido a dois quesitos, a dificuldade na atualização da base e a importância que foi dada ao levantamento teórico, e sua influência na tomada de decisões.

A revisão bibliográfica apresentou um número de fatores interessantes para avaliação, durando mais do que o esperado, porém dentro de uma faixa aceitável de tempo. A atualização do banco de dados no momento encontra-se praticamente finalizada, mesmo após alguns contratempos, que fizeram a tarefa durar mais do que o esperado, mesmo iniciando antes do previsto. Enquanto a geração do grafo a partir da base foi realizada de acordo com o previsto, mas a definição de características ainda está em fase de discussão. Essa geração permitiu a importação para um ambiente de visualização com sucesso, visto no capítulo anterior. Já a fase de implementação e avaliação inicia-se este mês e se dará como previsto, assim como as tarefas subsequentes, para as quais será mantido o cronograma.

1. Revisão bibliográfica;
2. Atualização da base MySQL;

3. Definição das características e geração do grafo a partir da base;

4. Importação para ambiente de visualização de grafos;

5. Implementação dos métodos de previsão e avaliação;

6. Comparação com as redes geradas;

7. Avaliação dos métodos;

8. Escrita do TCC.

Tabela 6.1: Cronograma

Semestre	2018 - 2					2019 - 1						
Etapa / Mês	AGO	SET	OUT	NOV	DEZ	JAN	FEV	MAR	ABR	MAI	JUN	JUL
1	X	X	X									
2		X	X	X								
3		X	X									
4			X	X								
5				X	X	X	X	X				
6								X	X	X		
7								X	X	X	X	
8		X	X	X	X	X	X	X	X	X	X	

Referências

ADAMIC, L. A.; ADAR, E. Friends and neighbors on the Web. *Social Networks*, v. 25, n. 3, p. 211–230, 2003. ISSN 03788733.

AL-OUFI, S.; KIM, H.-N.; El Saddik, A. Controlling privacy with trust-aware link prediction in online social networks. *Proceedings of the Third International Conference on Internet Multimedia Computing and Service - ICIMCS '11*, p. 86, 2011. Disponível em: <http://dl.acm.org/citation.cfm?doid=2043674.2043699>.

BALANCIERI, R. Análise de Redes de Pesquisa em uma Plataforma de Gestão em Ciência e Tecnologia: Uma Aplicação à Plataforma Lattes. . *Engenharia de Produção*, Mestrado, 2004. Disponível em: <http://teses.eps.ufsc.br/Resumo.asp?5621>.

BARABÁSI, A. L.; ALBERT, R. Emergence of scaling in random networks. *Science*, v. 286, n. 5439, p. 509–512, 1999. ISSN 00368075.

BARABÁSI, A. L. et al. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, v. 311, n. 3-4, p. 590–614, 2002. ISSN 03784371.

BARTAL, A.; SASSON, E.; RAVID, G. Predicting links in social networks using text mining and SNA. *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining, ASONAM 2009*, p. 131–136, 2009. ISSN 0192415X.

BASTIAN, M.; HEYMANN, S.; JACOMY, M. Gephi: An open source software for exploring and manipulating networks. 2009. Disponível em: <http://www.aaai.org/ocs-/index.php/ICWSM/09/paper/view/154>.

BECKER, G. S. A theory of social interactions. *Journal of Political Economy*, University of Chicago Press, v. 82, n. 6, p. 1063–1093, 1974. ISSN 00223808, 1537534X. Disponível em: <http://www.jstor.org/stable/1830662>.

BLISS, C. et al. An Evolutionary Algorithm Approach to Link Prediction in Dynamic Social Networks. *arXiv*, p. 22–26, 2013.

BOCCALETTI, S. et al. Complex networks: Structure and dynamics. *Physics Reports*, v. 424, n. 4-5, p. 175–308, 2006. ISSN 03701573.

BOWYER, K. W. et al. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813, 2011. Disponível em: <http://arxiv.org/abs/1106.1813>.

BRANDÃO, M. A. et al. Using link semantics to recommend collaborations in academic social networks. *Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion*, n. Section 3, p. 833–840, 2013. Disponível em: <http://dl.acm.org/citation.cfm?doid=2487788.2488058>.

DUNLAVY, D. M.; KOLDA, T. G.; ACAR, E. Temporal Link Prediction using Matrix and Tensor Factorizations. 2010. ISSN 15564681.

FAWCETT, T. Roc graphs: Notes and practical considerations for researchers. *Machine Learning*, v. 31, p. 1–38, 01 2004.

FENG, J. et al. Summarization-based mining bipartite graphs. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, p. 1249, 2012. ISSN 9781450314626. Disponível em: <http://dl.acm.org/citation-.cfm?doid=2339530.2339725>.

GAO, F. et al. Link Prediction Methods and Their Accuracy for Different Social Networks and Network Metrics. 2014. Disponível em: <http://eprints.bournemouth.ac-.uk/22934/1/[gamu15]link_prediction.pdf>.

GASPARINI, I.; KIMURA, M. H.; PIMENTA, M. S. Visualizando 15 anos de ihc. In: *Proceedings of the 12th Brazilian Symposium on Human Factors in Computing Systems*. Porto Alegre, Brazil, Brazil: Brazilian Computer Society, 2013. (IHC '13), p. 238–247. ISBN 978-85-7669-278-2. Disponível em: <http://dl.acm.org/citation.cfm?id=2577101-.2577149>.

GRYLLOS, P.; MAKRIS, C.; VIKATOS, P. Marketing campaign targeting using bridge extraction. *Proceedings of the Symposium on Applied Computing - SAC '17*, p. 1045–1052, 2017. Disponível em: <http://dl.acm.org/citation.cfm?doid=3019612.3019814>.

GUNS, R. Predictive Characteristics of Co-authorship Networks: Comparing the Unweighted, Weighted, and Bipartite Cases. *Journal of Data and Information Science*, v. 1, n. 3, p. 59–78, 2016. ISSN 2096-157X.

HAGBERG, A. A.; SCHULT, D. A.; SWART, P. J. Exploring network structure, dynamics, and function using networkx. In: VAROQUAUX, G.; VAUGHT, T.; MILLMAN, J. (Ed.). *Proceedings of the 7th Python in Science Conference*. Pasadena, CA USA: , 2008. p. 11 – 15.

HALL, M. et al. The WEKA data mining software: an update. *SIGKDD Explorations*, v. 11, n. 1, p. 10–18, 2009.

HASAN, M. A.; SALEM, S.; ZAKI, M. Link Prediction using Supervised Learning. *New York*, 2006. ISSN 10987576.

HASAN, M. A.; ZAKI, M. J. *Social Network Data Analytics*. 2011. ISBN 978-1-4419-8461-6. Disponível em: <http://link.springer.com/10.1007/978-1-4419-8462-3>.

HEIDER, F. *The Psychology of Interpersonal Relations*. Taylor & Francis, 2013. ISBN 9781134922253. Disponível em: <https://books.google.com.br/books?id=1XUzo-0gHvUC>.

HOSSIN, M.; SULAIMAN, M. A REVIEW ON EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATIONS. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, v. 5, n. 2, p. 1–11, 2015. ISSN 2231007X.

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*. New York, New York, USA: ACM Press, 2005. p. 141. ISBN 1581138768. Disponível em: <http://portal.acm.org/citation.cfm?doid=1065385.1065415>.

HUMPHRIES, M. D.; GURNEY, K. Network 'small-world-ness': A quantitative method for determining canonical network equivalence. *PLoS ONE*, v. 3, n. 4, 2008. ISSN 19326203.

JAMES, G. et al. *An Introduction to Statistical Learning: With Applications in R.* : Springer Publishing Company, Incorporated, 2014. ISBN 1461471370, 9781461471370.

JEH, G.; WIDOM, J. Simrank: A measure of structural-context similarity. ACM, New York, NY, USA, p. 538–543, 2002. Disponível em: <http://doi.acm.org/10.1145/775047-.775126>.

JONES, M. *Introduction to HCI*. 2016. Disponível em: <https://www.cs.bham.ac.uk-/~rxb/Teaching/HCI%20II/intro.html>.

JULIAN, K.; LU, W. Application of Machine Learning to Link Prediction. p. 3–7, 2015.

KATZ, J.; MARTIN, B. R. What is research collaboration? *Research Policy*, v. 26, n. 1, p. 1 – 18, 1997. ISSN 0048-7333. Disponível em: <http://www.sciencedirect.com-/science/article/pii/S0048733396009171>.

KIM, J.; WILHELM, T. What is a complex graph? *Physica A: Statistical Mechanics and its Applications*, v. 387, n. 11, p. 2637–2652, 2008. ISSN 03784371.

KOTSIANTIS, S. B. Mössbauer study of Fe-Re alloys prepared by mechanical alloying. *Hyperfine Interactions*, v. 237, n. 1, p. 1–8, 2007. ISSN 15729540.

LESKOVEC, J.; KLEINBERG, J.; FALOUTSOS, C. Graph Evolution: Densification and Shrinking Diameters. v. 1, n. 1, 2006. ISSN 15564681. Disponível em: <http:/-/arxiv.org/abs/physics/0603229>.

LI, E. Y.; LIAO, C. H.; YEN, H. R. Co-authorship networks and research impact: A social capital perspective. *Research Policy*, Elsevier B.V., v. 42, n. 9, p. 1515–1530, 2013. ISSN 00487333. Disponível em: <http://dx.doi.org/10.1016/j.respol.2013.06.012>.

LI, L. et al. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, v. 2, n. 4, p. 431–523, 2005. Disponível em: <https://doi.org/10.1080/15427951.2005.10129111>.

LI, Z.; FANG, X.; SHENG, O. R. L. A Survey of Link Recommendation for Social Networks: Methods, Theoretical Foundations, and Future Research Directions. *Ssrn*, v. 9, n. 1, 2015. ISSN 1556-5068.

LIBEN-NOWELL, D. An Algorithmic Approach to Social Networks. *Language*, p. 120, 2005. Disponível em: <http://www.cs.carleton.edu/faculty/dlibenno/papers/thesis-/thesis.pdf>.

LIBEN-NOWELL, D.; KLEINBERG, J. The Link Prediction Problem for Social Networks. *Proceedings of the Twelfth Annual ACM International Conference on Information and Knowledge Management (CIKM)*, n. November 2003, p. 556–559, 2003. ISSN 1532-2882.

LIBEN-NOWELL, D.; KLEINBERG, J. The Link Prediction Problem for Social Networks *. 2004. Disponível em: <www.arxiv.org.>.

LICHTENWALTER, R. N.; LUSSIER, J. T.; CHAWLA, N. V. New perspectives and methods in link prediction. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*, p. 243, 2010. Disponível em: <http://dl.acm.org/citation.cfm?doid=1835804.1835837>.

LINYUAN, L. L.; ZHOU, T. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, Elsevier B.V., v. 390, n. 6, p. 1150–1170, 2011. ISSN 03784371. Disponível em: <http://dx.doi.org/10.1016/j.physa.2010.11.027>.

LIU, W.; Lü, L. Link prediction based on local random walk. *EPL (Europhysics Letters)*, v. 89, n. 5, p. 58007, 2010. Disponível em: <http://stacks.iop.org/0295-5075/89/i=5-/a=58007>.

LÜ, L.; JIN, C.-H.; ZHOU, T. Similarity index based on local paths for link prediction of complex networks. *Phys. Rev. E*, American Physical Society, v. 80, p. 046122, Oct 2009. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevE.80.046122>.

LÜ, L.; JIN, C. H.; ZHOU, T. Similarity index based on local paths for link prediction of complex networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, v. 80, n. 4, p. 1–9, 2009. ISSN 15393755.

MARTÍNEZ, V.; BERZAL, F.; CUBERO, J.-C. A Survey of Link Prediction in Complex Networks. *ACM Computing Surveys*, v. 49, n. 4, p. 1–33, 2016. ISSN 03600300. Disponível em: <http://dl.acm.org/citation.cfm?doid=3022634.3012704>.

MARUYAMA, W. T.; DIGIAMPIETRI, L. A. Co-authorship prediction in academic social network. p. 79–90, 2016.

MCPHERSON, M.; SMITH-LOVIN, L.; COOK, J. M. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, v. 27, n. 1, p. 415–444, 2001. Disponível em: <https://doi.org/10.1146/annurev.soc.27.1.415>.

MEDINA, M. *Algoritmos e Programacao: Teoria e Pratica.* : Novatec, 2005. ISBN 857522073X.

MENDONÇA, F. *Geração das Redes de Colaboração Científi ca da Comunidade Acadêmica de IHC.* Tese (Trabalho de Conclusão de Curso) — Universidade do Estado de Santa Catarina - UDESC, 2017.

NARIN, F.; STEVENS, K.; WHITLOW, E. S. Scientific co-operation in europe and the citation of multinationally authored papers. *Scientometrics*, v. 21, n. 3, p. 313–323, Jul 1991. ISSN 1588-2861. Disponível em: <https://doi.org/10.1007/BF02093973>.

NEGI, S.; CHAUDHURY, S. Link Prediction in Heterogeneous Social Networks. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16*, p. 609–617, 2016. Disponível em: <http://dl.acm-.org/citation.cfm?doid=2983323.2983722>.

NEWMAN, M. The Structure and Function of Complex Networks. *Society*, v. 45, n. 2, p. 167–256, 2003.

NEWMAN, M. E. J. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, National Academy of Sciences, v. 98, n. 2, p. 404–409, 2001. ISSN 0027-8424. Disponível em: <http://www.pnas.org/content/98/2/404>.

PAPADIMITRIOU, A.; SYMEONIDIS, P.; MANOLOPOULOS, Y. Fast and accurate link prediction in social networking systems. *Journal of Systems and Software*, Elsevier Inc., v. 85, n. 9, p. 2119–2132, 2012. ISSN 01641212. Disponível em: <http://dx.doi.org/10.1016/j.jss.2012.04.019>.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

POWERS, D. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness and Correlation. n. December, 2007.

SAVAGE, J. et al. Chemical Reactant Recommendation Using a Network of Organic Chemistry. *Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17*, p. 210–214, 2017. Disponível em: <http://dl.acm.org/citation-.cfm?doid=3109859.3109895>.

SCHOLZ, M. *Node degree distribution*. 2018. Disponível em: <http://comissoes.sbc.org-.br/ce-ihc/documentos/da-importancia-dos-IHCs_2006.html.>

SOARES, P. R. da S.; PRUDêNCIO, R. B. C. Time Series Based Link Prediction. p. 10–15, 2012.

SOUZA, C. S. *Da importância dos Simposios Brasileiros de Fatores Humanos em Sistemas Computacionais*. 2006. Disponível em: <http://comissoes.sbc.org.br/ce-ihc-/documentos/da-importancia-dos-IHCs_2006.html.>

TONG, H.; FALOUTSOS, C.; PAN, J. Fast random walk with restart and its applications. p. 613–622, Dec 2006. ISSN 1550-4786.

WANG, P. et al. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences*, v. 58, n. 1, p. 1–38, 2014. ISSN 1674733X.

WATTS, D. J. J.; STROGATZ, S. H. H. Collective dynamics of 'small-world' networks. *Nature*, v. 393, n. 6684, p. 440–442, 1998. ISSN 0028-0836.

YANG, Y.; LICHTENWALTER, R. N.; CHAWLA, N. V. arXiv : 1505 . 04094v1 [cs . IR] 15 May 2015 Evaluating Link Prediction Methods. p. 1–35, 2015.

YANTAO, J. et al. Structural-interaction link prediction in microblogs. *Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion*, p. 193–194, 2013. Disponível em: <http://dl.acm.org/citation.cfm?doid=2487788.2487885>.

YIN, D.; HONG, L.; DAVISON, B. D. Structural link analysis and prediction in microblogs. *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, p. 1163, 2011. Disponível em: <http://dl.acm.org-/citation.cfm?doid=2063576.2063743>.

YU, Q. et al. Predicting Co-Author Relationship in Medical Co-Authorship Networks. *PLoS ONE*, v. 9, n. 7, p. 1–7, 2014. ISSN 19326203.

ZHU, J.; HONG, J.; HUGHES, J. Using Markov models for web site link prediction. *Thirteenth ACM Congerence on Hypertext and Hypermedia*, p. 169, 2002.