***Análise de Redes de Colaboração Científica:
Uma Abordagem Baseada em Grafos
Relacionais com Atributos***

Evelyn Perez Cervantes

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
COMO PARTE DOS REQUISITOS
PARA OBTENÇÃO DO GRAU DE
MESTRE EM CIÊNCIAS DA COMPUTAÇÃO
Programa: **Ciência da Computação**
Orientador: **Prof. Dr. Roberto M. Cesar-Junior**

Durante a elaboração deste trabalho, a autora recebeu auxílio financeiro da CAPES

- São Paulo, - 2015 -

Agradecimentos

Agradeço a Deus por todas as bênçãos, por me conceder uma família tão linda e por ter me trazido tão longe de casa para poder estudar no IME.

Meus sinceros agradecimentos aos professor Roberto Marcondes Cesar Junior e Jesús Mena Chalco, pela oportunidade de aprender com eles, pela paciência, a compreensão e pelo suporte nos momentos difíceis.

Agradeço à minha mãe, Catalina Cervantes, e minhas irmãs, Sidney P. Cervantes e Nataly P. Cervantes, por ficarem sempre ao meu lado, mesmo estando longe. Também porque sempre confiaram em mim e porque são minha maior motivação e fortaleza.

Agradeço à minha mãe brasileira, Jihan Zoghbi, porque foi um anjo em meu caminho; pela grande ajuda desinteressada quando mais precisei.

À CAPES pelo apoio financeiro, e a todos os funcionários do IME.

Finalmente, um agradecimento muito especial para o Wellington Goncalves, pela paciência, carinho e grande amor; porque sempre segurou minha mão e me transmitiu calma.

Resumo

PEREZ-CERVANTES, E. **Análise de Redes de Colaboração Científica: Uma Abordagem Baseada em Grafos Relacionais com Atributos**. 82 p. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2015.

A análise de redes sociais permite estudar a maneira como são estabelecidas as conexões entre indivíduos e como estas evoluem ao longo do tempo. A coautoria é uma das formas mais estudadas e bem documentadas de colaboração científica. Existem muitos aspectos de redes de colaboração científica, os quais podem ser rastreados de forma confiável através da análise de redes de colaboração usando métodos bibliométricos. Diversos esforços em diferentes áreas de pesquisa tentam analisar, entender, explicar e prever o comportamento de sistemas modelados através de redes sociais. Nestes estudos, os indivíduos são modelados como vértices de um grafo, enquanto as relações entre eles são representadas por arestas. Atualmente, o estudo de redes de colaboração científica é importante e necessário para apoiar o planejamento estratégico, implementação e gestão dos programas de pesquisa científica.

Neste trabalho, apresentamos um modelo de análise de redes científicas baseado em Grafos Relacionais com Atributos (ARG). O modelo proposto permite representar as redes de colaboração científica incluindo atributos individuais dos pesquisadores e atributos dos trabalhos colaborativos de pares de pesquisadores. Os dados correspondem às produções científicas de pesquisadores cadastrados na plataforma Lattes e extraídas automaticamente usando a ferramenta scriptLattes. Na primeira etapa, foi implementado o cálculo automatizado da taxa de internacionalização de cada pesquisador, a qual mostra a proporção entre o número de publicações internacionais e o número total de publicações. Esta medida junto com a produção científica individual discretizada em diversos grupos fazem parte das informações armazenadas nos vetores de atributos dos vértices dos ARGs. Por outro lado os vetores de atributos das arestas armazenam informações dos trabalhos colaborativos discretizados segundo a classificação da CAPES.

Adicionalmente, neste trabalho foram exploradas duas aplicações relacionadas à (i) predição de trabalhos colaborativos futuros e à (ii) influência dos pesquisadores na rede de colaboração. O resultado da predição de vínculos foi usado para determinar a influência dos pesquisadores na rede de colaboração. A influência tem sido explorada com base na variação da predição de ligações com a presença ou a ausência do pesquisador na rede.

Nossa proposta foi avaliada considerando diferentes testes sobre redes de coautoria científica de diversos grupos de pesquisadores. Os resultados obtidos são promissores para a análise de redes sociais em geral.

Palavras-chave: Análise de redes sociais, grafo relacional com atributos, redes de colaboração científica, predição de ligações, taxa de internacionalização, centralidade de vértices.

Abstract

The social network analysis allows the study of how the relationships are established between individuals and how their are evolving with the time. The co-authorship is one of the most studied and documented scientific collaboration. There are some aspects which could be traced in a reliable way through the social network analysis using bibliometric methods.

There are several proposals in different research areas trying to analyse, understand, explain and predict the behaviour of systems modeled as social networks. In this study, the individuals are modeled as vertices of a graph, while the relationships between them are represented by edges. Currently the study of scientific collaboration networks is important and necessary to support the strategic planning, implementation and management of scientific research programs.

In this work, we present an scientific networks analysis model based on Attributed Relational Graphs (ARG). The proposed model allows to represent the scientific collaboration networks including individual attributes of researchers and attributes of the collaborative work of researchers pairs. The data correspond to the scientific production of researchers, registered in the Lattes Platform and automatically extracted using the tool scriptLattes [Mena-Chalco e Cesar-Jr, 2009]. In the first step, was implemented the automated computation of the internationalization rate for each researcher, that shows the ratio between the number of international publications and the total number of publications.

This measure together with the individual scientific production discretized in diverse groups form part of the information stored in the vertices of the ARGs. On the other hand, the edges store information of collaborative work discretized according to the CAPES classification.

Additionally, this work explores two related applications (i) prediction of future collaborative work and (ii) influence of researchers in collaboration network. The result of the link prediction was used to determine the influence of researchers in collaborative networks. The influence in collaboration network is computed based on the variation of the link prediction with the presence or absence of the researcher in the network.

Our proposal was evaluated with different real scientific co-authorship networks and with different research groups. The results obtained look promising for analyzing social networks in general. **Keywords:** Social network analysis, attributed relational graph, scientific collaboration networks, link prediction, internationalization rate, vertex centrality.

Conteúdo

Resumo	iii
Abstract	iv
Lista de Figuras	vii
Lista de Tabelas	ix
1 Introdução	1
1.1 Motivação	2
1.2 Objetivos	3
1.3 Organização do Trabalho	3
2 Revisão bibliográfica	4
2.1 Propriedades das redes sociais	5
2.2 Modelos de redes	8
2.3 Grafo relacional com atributos	10
2.4 Análise da taxa de internacionalização	12
2.5 Predição de ligações	15
2.6 Influência na rede de colaboração	18
3 Modelagem de redes usando ARGs	22
3.1 Método proposto	22
3.2 Análise da taxa de internacionalização	26
3.2.1 Descrição dos dados	29
3.2.2 Resultados	29
3.3 Predição de ligações	35
3.3.1 Modelo de classificação	39
3.3.2 Descrição dos dados	40
3.3.3 Resultados	50
3.4 Influência na rede de colaboração	51
3.4.1 Descrição dos dados	53
3.4.2 Resultados	54

4	Considerações Finais	63
4.0.3	Trabalhos futuros	64
4.0.4	Publicacoes	65
	Bibliografia	66

Lista de Figuras

2.1	Grafo ou rede	5
2.2	Distribuição de graus de uma rede de colaboração	6
2.3	Assortatividade de uma rede de colaboração real	7
2.4	Representação gráfica de um grafo relacional com atributos (ARG)	12
2.5	Grafo relacional com atributos com seus respectivos BARGs	12
2.6	Representação gráfica do problema de predição de ligações	15
2.7	Principais modelos propostos para predição de ligações	16
3.1	Grafo relacional com atributos (ARG)	22
3.2	Esquema de alto nível da proposta	23
3.3	Arquivo padrão para armazenar ARGs	24
3.4	ARG com atributos nos vértices e arestas	25
3.5	Processo para colapsar vários ARGs	25
3.6	Procedimento para o cálculo do IPR	27
3.7	Medidas quantitativas acadêmicas para as áreas de conhecimento	31
3.8	Mapas mundiais de colaborações acadêmicas para as diferentes áreas	34
3.9	Representação gráfica do procedimento de construção do preditor de ligações	35
3.10	Representação gráfica do procedimento de construção do preditor de ligações	35
3.11	Representação gráfica do procedimento da criação dos vetores de características	40
3.12	Representação gráfica do procedimento de adaptação do algoritmo de aprendizado	40
3.13	Rede de colaboração da astronomia	42
3.14	Rede de colaboração da ciência da computação	43
3.15	Rede de colaboração da física	44
3.16	Rede de colaboração da geociências	45
3.17	Rede de colaboração da matemática	46
3.18	Rede de colaboração da oceanografia	47
3.19	Rede de colaboração da probabilidade	48
3.20	Rede de colaboração da química	49
3.21	Desempenho individual das características consideradas no vetor do preditor	50
3.22	Desempenho individual das características consideradas no vetor do preditor	51
3.23	Desempenho em grupo das características consideradas no vetor do preditor	52
3.24	Representação gráfica do método proposto para detectar a influência de colaboração	53
3.25	Representação gráfica do processo de criação do classificador	54
3.26	Dinâmica das arestas na rede de colaboração dos programas de pós-graduação em Ciência da Computação no Brasil	55

3.27 Rede de colaboração de todos os pesquisadores afiliados a 45 programas de pós-graduação em Ciência Computação no Brasil	56
3.28 Ranking da influência colaborativa com sinal e sem sinal	58
3.29 Correlação entre as métricas de centralidade conhecidas e a medida proposta	59
3.30 Rede de colaboração de pesquisadores no Brasil	61
3.31 Ranking da influência colaborativa em redes ponderadas	62

Lista de Tabelas

3.1	Quantidade de pesquisadores para as áreas de conhecimento consideradas.	30
3.2	Resumo dos resultados obtidos para os grupos brasileiros de pesquisa, associados com cinco grandes áreas do conhecimento. P indica número de publicações, P-DOI indica número de publicações com DOI registrado, P-I indica número de publicações internacionais. O valor entre parênteses indica a proporção em relação a P-DOI.	30
3.3	O <i>top</i> dos 30 países que mais colaboram com o Brasil. A quantidade numérica indica o número de publicações.	33
3.4	Quantidade de pesquisadores para as áreas de conhecimento consideradas.	41
3.5	Média dos valores obtidos da Acurácia, Precisão, Sensibilidade e Medida-F do preditor de ligação.	50

Capítulo 1

Introdução

Devido à existência de uma grande variedade de sistemas que podem ser modelados como redes, o desenvolvimento de projetos em diferentes áreas de pesquisa tem sido incrementado. O principal objetivo desses projetos é analisar, entender, explicar e prever o comportamento de tais sistemas através de redes [Al Hasan e Zaki, 2011; Getoor e Diehl, 2005; Lü e Zhou, 2011]. A concepção de rede abarca teorias, modelos e aplicações que são expressos em termos de conceitos relacionais. No caso particular das redes sociais, a unidade de estudo não é o indivíduo, mas sim uma entidade composta por um conjunto de indivíduos e as ligações entre eles [Wasserman, 1994].

A compreensão das estruturas sociais das redes tem implicações importantes em muitos aspectos, incluindo provisionamento, *bootstrapping*, identificação de padrões de comportamento, identificação de indivíduos influentes e suas ligações. As aplicações são realizadas em diferentes áreas como: esporte, *marketing*, ciências sociais, psicologia, dentre outros. Estudos tradicionais em redes sociais estão focados na observação da estrutura da rede. Nestes modelos os indivíduos são representados por vértices e as relações entre eles são representadas por arestas. Resultados recentes sugerem que o incremento de atributos dos indivíduos (localização, empregador, por exemplo) na estrutura da rede pode fornecer uma compreensão mais refinada das redes sociais [Gong *et al.*, 2011; Yang *et al.*, 2011; Zhou *et al.*, 2009b]. Além disso, o uso de atributos sociais pode esclarecer as propriedades estruturais da rede, reconhecer atributos estruturais e determinar como esses atributos influenciam sua estrutura. No entanto, existem poucos estudos para fornecer uma compreensão sistemática desses efeitos nas redes sociais [Zhenqiang G *et al.*, 2012].

A crescente complexidade do desenvolvimento científico e tecnológico junto com as mudanças das políticas de educação criaram a necessidade de medir a atividade dos diferentes grupos de pesquisadores, frequentemente modelada como redes de colaboração científica. Em geral, a ciência e tecnologia apresentam diferentes dimensões que podem ser mensuradas por meio de indicadores qualitativos e quantitativos [Neto, 2010]. Atualmente, existe um interesse especial por parte das entidades ou órgãos de fomento em fazer estudos de avaliação que permitam (i) analisar a infraestrutura disponível e políticas de investimento em pesquisa científica e tecnológica; (ii) analisar

as tendências e variáveis do setor, permitindo a formulação de políticas científicas e (iii) elaborar estratégias e planos de ações a curto e longo prazo.

Neste trabalho, estudamos redes de colaboração científica nas quais dois cientistas/pesquisadores, representados por vértices, são considerados conectados caso tenham algum tipo de colaboração em conjunto. Dada a disponibilidade de informações não usadas atualmente na análise de tais redes, é proposto o uso de Grafos Relacionais com Atributos (ARG) para integrar a estrutura da rede e os novos atributos de uma forma escalável e obter uma análise mais refinada das redes científicas. Este trabalho foi realizado sob a coorientação do Professor Jesús P. Mena-Chalco da Universidade Federal do ABC.

1.1 Motivação

- As redes de colaboração científica reais dificilmente têm vínculos simples. Ao invés disso é comum encontrar redes científicas multirelacionais, isto é, redes de pesquisadores que mantêm vínculos de diferentes tipos (coautoria, orientação, etc.); estes tipos de redes são chamadas de *redes de informação heterogênea* [Han, 2009; Wasserman, 1998], termo genérico que encapsula redes *multi-modo*, *multi-dimensional* e *multirelacional*. Alguns dos problemas cruciais na análise de redes científicas multirelacionais estão vinculados à qualidade dos dados e à capacidade de lidar com os diferentes tipos de vínculos entre os pesquisadores. Existem também bancos de dados, como a plataforma Lattes, ricos em atributos descritivos dos pesquisadores que são pouco aproveitados. Nos últimos anos, tem se dado uma grande ênfase na análise da interação de redes científicas. No entanto, a natureza multi-dimensional dessas interações foi largamente ignorada, principalmente por causa da falta de dados ou de um modelo que dê o suporte necessário.
- Objetivos diferentes requerem dados diferentes, além de uma análise diferente. No entanto, a organização dos dados não deveria depender do tipo de análise que será feita, de modo que a melhor forma de organizar os dados é centralizar e consolidar o armazenamento em uma única estrutura suficientemente flexível para produzir os dados precisos quando for necessário. Há também casos nos quais é preciso obter resultados específicos de modo que os dados não podem ser considerados separadamente. Essa estrutura deve permitir armazenar os diferentes atributos tanto dos pesquisadores como das relações entre eles, além de sustentar a relevância dos atributos. A estrutura que cumpre os requerimentos antes mencionados é chamada de Grafo Relacional com Atributos (ARG) [Tsai e Fu, 1979].

1.2 Objetivos

O objetivo geral desse trabalho de mestrado é desenvolver e propor um modelo de análise de redes de colaboração científica baseado em Grafos Relacionais com Atributos.

1. Adaptar e desenvolver um método de criação de redes de colaboração científica;
2. Enriquecer as redes de colaboração científica com dados extraídos da plataforma Lattes;
3. Definir e implementar atributos descritivos e estruturais associados aos elementos das redes de colaboração científica;
4. Representação dos grafos relacionais com atributos através de arquivos;
5. Avaliação do modelo proposto através das aplicações de predição de ligações, análise de taxa de internacionalização e influência colaborativa. Finalmente, aplicar o método a redes reais com a finalidade de gerar conhecimento útil na tomada de decisões.

1.3 Organização do Trabalho

O texto está organizado da seguinte forma: o Capítulo 2 é dedicado à revisão da literatura sobre trabalhos relacionados a este projeto, são descritas algumas características que tornam especiais as redes sociais e também são apresentados os grafos relacionais com atributos. No Capítulo 3, é descrito o modelo proposto, juntamente com a descrição detalhada das aplicações escolhidas para avaliação de nossa proposta. Nesse mesmo capítulo são, apresentados alguns dos resultados obtidos para cada uma das aplicações. Por fim, o Capítulo 4 encerra o documento, apresentando as considerações e conclusões finais. Nesse capítulo também são apresentadas algumas sugestões de trabalhos futuros.

Capítulo 2

Revisão bibliográfica

Neste capítulo, a análise de redes sociais é apresentada como uma perspectiva de pesquisa distinta no âmbito das ciências sociais e comportamentais; pois está baseada na hipótese da importância das relações entre as unidades que interagem. A perspectiva de rede social engloba teorias, modelos e aplicações que são expressas em termos de conceitos ou processos relacionais. Com o crescente interesse e aumento da utilização da análise de redes, chegou-se em um consenso sobre os princípios centrais e, somado à utilização dos conceitos relacionais, são considerados importantes aspectos tais como: Os atores e suas ações são vistos como interdependentes em vez de unidades autônomas e independentes, os vínculos relacionais entre atores são canais para a transferência ou “fluxo” de recursos, os modelos de rede com foco em indivíduos veem o ambiente estrutural da rede como oportunidades para fornecer restrições sobre a ação individual; os modelos de rede conceituam estrutura (social, econômica, política e assim por diante) como padrões duradouros de relações entre os atores.

A unidade para análise de rede não é o indivíduo, mas uma entidade composta por um conjunto de indivíduos e as ligações entre eles. Métodos de rede se concentram em duplas (dois atores e seus laços), tríades (três atores e seus laços) ou sistemas maiores (subgrupos de indivíduos, ou redes inteiras) [Wasserman, 1998]. Neste contexto, a análise de redes fornece um vocabulário e um conjunto de definições formais para expressar conceitos teóricos e propriedades; alguns dos mais importantes serão apresentados nas seções seguintes. Adicionalmente, serão apresentados alguns modelos de redes que permitem testar e avaliar os conceitos subjacentes à perspectiva de rede.

Uma rede ou grafo $G = (V, E)$ é uma estrutura que consiste em um conjunto finito não vazio de vértices $V = V(G)$ e um conjunto $E = E(G)$ formado por pares de elementos de V , chamados de arestas de G [Bondy e Murty, 1976]. A Figura 2.1 mostra um grafo $G = (V, E)$ com cinco vértices e oito arestas em que $V(G) = v_1, v_2, v_3, v_4, v_5$ e $E(G) = e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8$.

Uma maneira alternativa de representar um grafo é utilizando matrizes. Seja um grafo $G = (V, E)$ com n vértices; a matriz de adjacência de G , $A(G)$, é uma matriz de ordem n cujos elementos

Figura 2.1: Grafo ou rede

são:

$$a_{ij} = \begin{cases} 1 & \text{se } \{v_i, v_j\} \in E \\ 0 & \text{caso contrário} \end{cases}$$

Portanto, a matriz de adjacência do grafo apresentado na Figura 2.1 é a seguinte:

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

Nos modelos de grafos, os atores são representados por vértices e as relações entre eles são representadas por arestas; comumente, são considerados importantes atores envolvidos em várias relações.

2.1 Propriedades das redes sociais

A seguir são apresentadas algumas das características que são exploradas neste trabalho e que por sua vez mostram que as redes sociais não cumprem algumas características típicas de outros tipos de redes como as aleatórias.

Distribuição dos graus

Seja uma rede $G = (V, E)$ e p_k a probabilidade de um vértice de grau k seja escolhido aleatoriamente, em que $1 \leq k \leq N_p$. A distribuição de tais probabilidades é chamada de distribuição de grau p_k da rede N .

O grau de um vértice é o número de arestas adjacentes a ele. Normalmente, os vértices de redes não sociais apresentam distribuições Gaussianas ou de *Poisson*, no caso de ser aleatória. As redes sociais, por sua parte, têm outros tipos de distribuições como a distribuição *Power Law* [Adamic e

Figura 2.2: Distribuição de graus de uma rede de colaboração real sendo possível observar que existem poucos vértices com muitas ligações, enquanto existem muitos vértices com poucas ligações.

Adar, 2003], segundo esta distribuição, existem poucos indivíduos (vértices) com graus altos e um grande número de vértices com graus baixos.

A Figura 2.2 mostra uma rede de colaboração científica real, em que os vértices representam pesquisadores ligados por arestas representando o trabalho de colaboração. Nela é possível observar que existem poucos vértices que atraem muitas colaborações, os vértices maiores e pretos. No outro extremo, existem muitos vértices com poucas colaborações, vértices menores.

Assortatividade

Essa propriedade segue o princípio pelo qual tendemos a ser semelhantes aos nossos amigos. Vistos coletivamente, um grupo de amigos tem características semelhantes. Assortatividade pode ser computada usando alguma medida específica de similaridade. Na teoria de grafos, uma característica comumente usada é o grau dos vértices. Nesse caso, uma rede com alto grau de assortatividade irá conter arestas entre vértices com graus semelhantes. Em geral, se o valor da assortatividade $r \in [-1, 1]$ é maior do que zero, a rede contém arestas entre vértices semelhantes; se o valor é menor do que zero, existem arestas entre vértices muito diferentes em relação à medida de semelhança que está sendo avaliada. [Newman, 2003; Newman e Park, 2003].

O coeficiente de assortatividade é calculado da seguinte forma:

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}. \quad (2.1)$$

em que e_{ij} é a fração de arestas da rede que conecta um vértice do tipo i com um vértice do tipo j . Esta fórmula devolve $r = 0$ se a rede for não assortativa e $r = 1$ quando é perfeitamente assortativa.

A Figura 2.3 mostra uma rede de colaboração científica real com índice de assortatividade igual a 0,798. Neste gráfico, são mostrados em cores os grupos contendo vértices com características similares.

Figura 2.3: Rede de colaboração real com índice de assortatividade igual a 0,798. Cada cor representa um grupo com tipos de publicações semelhantes

Transitividade ou agrupamento

Transitividade é a probabilidade de um vértice ser ligado ao vizinho do seu vizinho. Essa propriedade quantifica a característica pela qual, em muitas redes sociais, se um vértice A está ligado ao vértice B e o vértice B está ligado a C , então existe uma probabilidade elevada de que o vértice A também será ligado ao vértice C . Socialmente falando, é provável que os amigos de seus amigos sejam seus amigos. Baseia-se no número relativo de triângulos no grafo, isto é, um sub-grafo plano com 3 vértices e 3 arestas, em relação ao número total de triplas de vértices conectados.

$$IA = \frac{3 \times \text{número de triângulos na rede}}{\text{número de triplas de vértices conectados}}. \quad (2.2)$$

O índice de agrupamento (IA) da rede de colaboração mostrada na Figura 2.2 é igual a 0,389.

Efeito mundo pequeno

Milgram, em 1960, fez um dos experimentos mais famosos na área de análise de redes sociais, mostrou que, passando cartas de pessoa em pessoa era possível chegar a um indivíduo alvo designado em apenas um pequeno número de passos, por volta de seis.

Para uma rede não dirigida, o cálculo da média dos caminhos mais curtos está dado pela seguinte fórmula:

$$l = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}, \quad (2.3)$$

em que d_{ij} é a distância mais curta entre os vértices i e j , e a distância de um vértice para si mesmo é zero.

Esta medida tem problemas quando a rede tem mais do que uma componente conexa, porque nesses casos irão existir pares de vértice sem caminho entre si. Por convenção, a ausência de caminho entre dois vértices é representado pelo valor infinito, portanto, o valor de l vai se tornar infinito. Para evitar esse problema, a fórmula é modificada para que seja uma média harmônica das distância entre pares de vértices. Isto é, a recíproca da média dos recíprocos, como visto na Equação (2.4).

$$l^{-1} = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}^{-1}. \quad (2.4)$$

Portanto, o valor infinito não contribui na soma.

A média dos caminhos mais curtos para a rede real de colaboração apresentada na Figura 2.2 é de 6,618.

Nesta seção foram apresentadas algumas das características que mostram que as redes sociais são especiais. Ao mesmo tempo foram avaliadas algumas características uma rede real de colaboração científica.

2.2 Modelos de redes

Na literatura existem diversos modelos para descrever sistemas de interação. Dado um grafo, este pode ser classificado dependendo do modelo gerador. Existem modelos padrão de geração e crescimento de grafos que tentam refletir algumas das características observadas nas redes reais, estes modelos contemplam os seguintes elementos chave:

1. Agregação contínua de vértices e arestas no tempo;
2. Ligação preferencial, isto é, existe maior probabilidade de uma aresta estar conexa a um vértice de grau alto do que a um de baixo grau;
3. Remoção ou movimento de arestas na rede no tempo;

A seguir, apresentamos alguns dos modelos geradores mais conhecidos:

Modelo de grafo aleatório Erdős-Rényi-Gilbert

Seja $G = (N, p)$ um grafo não dirigido com N vértices, no qual; cada aresta tem uma probabilidade p de existir. O modelo tem uma probabilidade binomial, sendo que a probabilidade de ter E arestas

é definido por:

$$l(G(N, p) \text{ ter } E \text{ arestas} | p) = p^E (1 - p)^{\binom{N}{2} - E} \quad (2.5)$$

Este é o modelo mais comumente adotado na literatura. Ele especifica a probabilidade p de cada aresta e controla o número esperado de arestas $p\binom{N}{2}$.

Outro modelo semelhante para um grafo $G = (N, E)$ não dirigido com N vértice e um número fixo de arestas E , escolhidas aleatoriamente dentre os $\binom{N}{2}$ possíveis, considera que cada aresta tem a mesma probabilidade de ser escolhida. Uma interpretação equivalente é que todos os grafos com N vértices e exatamente E arestas são igualmente prováveis e um deles é escolhido aleatoriamente.

Modelo aleatório generalizado

Existem muitas propostas que generalizam o modelo Erdős-Rényi-Gilbert. Em particular, uma das mais simples é incluir uma distribuição de graus não Poisson. Grafos aleatórios com uma distribuição de grau arbitrária $P(k)$ foram propostos na literatura. O modelo de configuração [Bornholdt e Schuster, 2006], permite amostrar grafos com uma sequência de graus definida. A sequência de graus é qualquer sequência de N números inteiros $D = k_1, k_2, \dots, k_N$ tal que $\sum_i k_i = 2|E|$, em que $|E|$ é o número de *links* no grafo. Neste modelo de configuração, D é escolhido de tal maneira que a fração dos vértices com grau k tenderá à distribuição de graus desejada $P(k)$, para valores grandes de N .

Esse modelo considera que do conjunto de todos os grafos com N vértices e uma sequência de grau determinado D , cada grafo tem a mesma probabilidade. A sequência de graus fixa D pode ser gerada aleatoriamente também atribuindo a cada vértice i um número médio de arestas equivalente ao valor esperado de k_i .

Esses dois modelos descritos acima não conseguem caracterizar a redes sociais. Portanto, foram propostos novos modelos que permitem descrever de maneira mais realista a dinâmica das redes sociais. A seguir, apresentamos dois dos modelos representativos.

Modelo de grafo de mundo pequeno

O modelo de Watts e Strogatz [Watts e Strogatz, 1998] incorpora tanto a propriedade do mundo pequeno quanto um alto coeficiente de agrupamento. Segundo este modelo, muitas redes apresentam padrões altamente conectadas, tendendo a formar pequenas quantidades de conexões em cada vértice.

O modelo baseia-se em um processo de religação das extremidades implementadas com uma probabilidade p . No início é criado um anel de N vértices, no qual cada vértice está simetricamente ligado a seus $2m$ vizinhos mais próximos, tendo um total $K = mN$ arestas. Então, o modelo mundo pequeno é criado tomando uma pequena fração das arestas neste grafo e “religando-as”. O

procedimento de religação se inicia passando por cada aresta e movendo uma extremidade da aresta escolhida para um outro vértice escolhido uniformemente ao acaso na rede com uma probabilidade p . Entretanto não são criadas aresta duplas ou ligações de um vértice para ele mesmo.

Modelo de grafo livre de escala

Existem redes que apresentam uma ordem em sua dinâmica, com características bem específicas. Uma das principais características, a denominada conexão preferencial, é a tendência de um novo vértice se conectar a um vértice da rede que tem um grau elevado de conexões [Barabási e Albert, 1999]. Essa característica implica em redes com poucos vértices altamente conectados, denominados *hubs*, e muito vértices com poucas conexões. As redes com essas características são denominadas livres de escala. A distribuição de graus dos vértices dessas redes não possui um valor característico e segue a lei de potência, isto é, nessas redes existem muitos vértices com poucas ligações e poucos vértices com muitas ligações. Portanto, nessas redes cumpre-se que vértices com graus altos tendem a se ligar entre eles e atrair novas ligações.

2.3 Grafo relacional com atributos

Com o avanço dos anos, a quantidade de informações armazenada e disponível nos bancos de dados está em constante crescimento, o que evidencia a necessidade de mudar ou fazer mais extensível e flexíveis as estruturas tradicionalmente usadas para representar os diversos tipos de interação social. A teoria de grafos, em geral, tem sido bastante utilizada tanto na representação quanto na análise de redes sociais. Neste sentido, a análise evoluiu a partir de propostas iniciais usando grafos simples binários, representando apenas a presença ou ausência de uma relação [Girvan e Newman, 2002; Newman e Park, 2003], depois foi incorporada a direcionalidade utilizando grafos dirigidos para representar a influência de um vértice sobre outro [Newman, 2003; Newman e Park, 2003; Sun *et al.*, 2011]. Posteriormente, foi acrescentada a importância da ligação utilizando grafos com pesos ou ponderados [Lü e Zhou, 2010; Yang *et al.*, 2012]. Trabalhos mais recentes usam multigrafos para representar múltiplas relações entre vértices [Ströele *et al.*, 2013; Szell *et al.*, 2010]. Estas propostas apresentam limitações baseadas no uso de informação incompleta.

Resultados recentes [Gong *et al.*, 2011; Yang *et al.*, 2011; Zheleva *et al.*, 2010; Zhou *et al.*, 2009b] sugerem que aumentar informações sociais dos indivíduo da rede (por exemplo, localização, empregador) pode fornecer uma compreensão mais refinada das redes sociais.

Existem vários problemas em diferentes áreas de pesquisa que podem ser modelados como grafos, que são largamente conhecidos e estudados. Este é o caso de grafos relacionais com atributos que são comumente usados para descrever objetos estruturados [Sanfeliu e King-Sun, 1983].

Os pioneiros no uso dessa estrutura foram [Tsai e Fu, 1979]. Este tipo de grafos é considerado uma ferramenta poderosa para descrever imagens e cenas no reconhecimento de padrões. Nessa estrutura, os vértices representam primitivas ou subpadrões de objetos estruturados e as arestas representam algum tipo de relação entre primitivas ou subpadrões. Neste contexto, para reconhecer um padrão, os dados são representados como um grafo relacional com atributos e é estabelecida uma função de correspondência com um padrão protótipo. Neste processo de geração de correspondência podem ser usadas técnicas de isomorfismo de grafos ou subgrafos, por exemplo.

Na área de visão computacional e processamento de imagens, um ARG basicamente pode ser definido como um grafo dirigido, no qual as informações de aparência são representadas por atributos de vértices, e as informações estruturais como atributos de arestas. Uma definição mais formal é feita a seguir:

Definição:. *um ARG (abreviatura em inglês de Attributed Relational Graphs), é um grafo dirigido no qual, um vetor de atributos é associado aos vértices e às arestas. Matematicamente, um ARG é definido da seguinte forma:*

$$G = (V, E, \mu, \nu)$$

No qual:

Onde:

- V é um conjunto finito não vazio de vértices.
- $E \subseteq V \times V$ é um conjunto finito não vazio de arestas.
- μ representa o conjunto finito e não vazio de vetores de atributos dos vértices.
- ν representa o conjunto finito e não vazio de vetores de atributos das arestas.

Segundo este modelos, será usada a seguinte convenção para denotar vértices e arestas:

Um vértice $v_i \in V$ será denotado pelo par (s_i, x_i)

Em que s_i é o símbolo que representa a estrutura do vértice v_i e $x_i = [x_{i1}, x_{i2}, \dots, x_{im}]$ denota o vetor com m atributos numéricos ou lógicos de v_i .

Uma aresta arbitrária entre os vértices v_i e v_j será denotada pelo par $e_k = (u_k, y_k)$, em que u é o símbolo que denota a estrutura de e_k , enquanto $y_k = [y_{k1}, y_{k2}, \dots, y_{kn}]$ denota um vetor com n atributos numéricos ou lógicos.

A Figura 2.4 mostra uma possível representação gráfica de um grafo relacional com atributos. Uma das formas mais conhecidas de o determinar o isomorfismo entre grafos relacionais com atributos correspondentes é decompô-los em subgrafos simples, BARG (abreviatura em inglês de Basic Attributed Relational Graphs), e encontrar a correspondência entre eles. Um BARG é um

Figura 2.4: Representação gráfica de um grafo relacional com atributos (ARG).

tipo de árvore de um nível, composta por um nó arbitrário e todos os nós conectados a ele. A Figura 2.5 exibe um exemplo de decomposição de um ARG em seus BARGs.

Figura 2.5: Grafo relacional com atributos com seus respectivos BARGs (adaptado de [Kalviainen e Oja, 1990]).

Nas seções anteriores, foram apresentados os conceitos básicos e revisão bibliográfica sobre os fundamentos para modelagem de redes sociais científicas usando grafos. Nas seções restantes do capítulo, serão tratadas aplicações importantes do modelo proposto na dissertação.

2.4 Análise da taxa de internacionalização

No Brasil, a internacionalização da ciência é atualmente uma questão-chave na agenda da política, tecnologia e inovação. Portanto, atividades voltadas para a internacionalização da investigação e desenvolvimento têm recebido especial interesse das agências de fomento e entidades que guiam as políticas de ciência e tecnologia no Brasil. Um exemplo nesse sentido é o programa “Ciência Sem Fronteiras”¹, criado em 2011.

Devido à globalização, a educação internacional tomou uma posição central na pesquisa em educação em todo o mundo [Dolby e Rahman, 2008]. A dimensão internacional do ensino superior está em constante mudança, e se tornando complexa e confusa [Knight, 2004]. Portanto, é importante (i) examinar a noção de internacionalização da ciência à luz das mudanças atuais,

¹www.cienciasemfronteiras.cnpq.br/web/english/theprogram

e (ii) criar mecanismos para medir quantitativamente a IPR (do inglês *International Publication Ratio*) de grupos de pesquisa. A colaboração internacional de atores acadêmicos (pesquisadores e instituições) pode ser encontrada de várias formas, tais como: intercâmbios (mobilidade) dos pesquisadores, reuniões de trabalho, organização de conferências internacionais, cooperação em projetos acadêmicos, entre outros. Todos esses tipos de colaboração podem produzir publicações em coautoria internacional como resultado da interação entre os atores acadêmicos envolvidos. É importante notar que a análise de trabalhos colaborativos (publicações) não fornece uma representação completa da colaboração científica internacional, mas dá uma representação interessante de alguns aspectos da colaboração internacional [Raan, 1998].

O mundo está passando por um processo de integração internacional. Sendo assim, a importância da colaboração científica internacional tem aumentado. Durante os últimos anos surgiram estratégias para aumentar a visibilidade internacional acadêmica, a qual envolve muitos estudos sobre internacionalização. Alguns deles se concentram na elaboração de estratégias em termos de relações científicas com o pesquisador ou instituições de países estrangeiros, enquanto outros tentam explicar o impacto da colaboração científica internacional em resultados de investigação com base em citações bibliográficas. Ambas as estratégias apresentaram conclusões interessantes [Hassan e Haddawy, 2012; Katz e Hicks, 1997; Knight, 2004; Raan, 1998; Zhuge, 2006]. A partir de estudos anteriores podem ser citadas: as ciências fundamentais tendem a apresentar um maior grau de colaboração internacional, quanto maior for o sistema nacional de pesquisa, menor é a colaboração internacional, por outro lado, publicações de autores individuais parecem ser menos citadas e têm menor impacto, ou seja, o maior impacto vem de publicações envolvendo colaboração (coautores) com instituições estrangeiras [Raan, 1998].

Nesta dissertação, uma publicação é chamada “internacional” se pelo menos um de seus coautores tem cadastrado seu vínculo com uma instituição estrangeira. Em outras palavras, uma publicação é considerada internacional se pelo menos um dos endereços dos coautores contém o nome de um país diferente do Brasil. O endereço dos autores foram extraídos diretamente das páginas *web* das bibliotecas digitais. Neste trabalho foi usada a medida IPR [Leite *et al.*, 2011], que representa a relação entre o número de publicações internacionais e o número total de publicações. Neste estudo só foram considerados os artigos completos publicados em periódicos; o cálculo foi feito automaticamente pelo módulo de internacionalização do *scriptLattes* [Perez-Cervantes *et al.*, 2012], o qual toma como base o endereço cadastrado pelos coautores, que representa o endereço das instituições às quais os coautores foram associados no momento da publicação.

A utilização de dados bibliométricos para estudar colaboração internacional não é um conceito novo. Uma análise da colaboração de pesquisadores coreanos utilizando o método bibliométrico sobre o número de países participantes revela que as publicações escritas por pesquisadores de três ou mais países diferentes têm aumentado consideravelmente nas últimas décadas na Coreia [Kim, 2006]. Neste trabalho é feita uma classificação das colaborações em duas categorias: *simétrica* e *assimétrica*. O primeiro tipo refere-se ao caso em que pesquisadores de diferentes países partici-

pam do projeto de pesquisa de maneira equivalente, enquanto o último está relacionado ao caso em que um subgrupo de pesquisadores representa a transferência de conhecimentos entre coautores. Em alguns países que têm experimentado o desenvolvimento científico nas últimas décadas, pode-se esperar um aumento do tipo simétrico de colaboração de pesquisa e uma diminuição da colaboração de pesquisa assimétrica [Kim, 2006].

Existem alguns esforços que tentam explicar a evolução da ciência brasileira tanto no contexto mundial quanto na América Latina, usando dados bibliográficos extraídos do ISI (do inglês *Institute for Scientific Information*). Segundo Glänzel e seus colaboradores [Glänzel *et al.*, 2006], o Brasil é líder na região e sua comunidade científica é caracterizada pela intensa colaboração internacional com o seu ambiente geopolítico, bem como com a América do Norte e Europa. No entanto, os pesquisadores preferem publicar em revistas nacionais ou regionais, e, às vezes, até mesmo em sua língua nativa. Este estudo mostra uma evolução da colaboração internacional do contexto regional para o contexto mundial, mostrando os Estados Unidos como um dos parceiros mais importantes fora da América Latina. Por outro lado, existe evolução da ciência em algumas regiões do país, mas é incipiente em outras. Os autores mostram que a Física, a Agricultura e a Biologia são os campos mais produtivos em colaboração internacional [Leta *et al.*, 2006].

Os efeitos naturais do crescimento da colaboração científica internacional são múltiplos. O volume brasileiro de colaboração internacional tem aumentado significativamente desde 1981. Estudos detalhados sugeriram que esse crescimento se deve ao aumento da participação brasileira no banco de dados do ISI. Publicações brasileiras vêm ganhando reconhecimento, através de citações e a produtividade científica tem crescido, não só com base em citações, mas também em cooperação direta entre os pesquisadores [Leta e Chaimovich, 2002]. Nesse contexto, o uso de mecanismos que permitem explorar e medir quantitativamente a IPR é cada vez mais necessário e útil para comparar vários grupos brasileiros de pesquisa. É interessante notar que nosso trabalho é baseado no DOI (do inglês *Digital Object Identifier*) associado a cada publicação registrada no currículo Lattes. O DOI é utilizado (i) como identificador da publicação e (ii) como um “link” para acessar o conteúdo digital na Internet. O DOI é representado por uma sequência alfanumérica exclusiva atribuída pela *International DOI Foundation* (IDF) para identificar o conteúdo e fornecer um link persistente à sua localização na Internet. Normalmente, ele é usado para identificar entidades físicas, digitais ou abstratas. A editora científica atribui um DOI específico quando um artigo é publicado e disponibilizado eletronicamente. Todos os números DOI começam com o algarismo 10 e contêm um prefixo e um sufixo separado por uma barra. O prefixo é um número único de quatro ou mais dígitos atribuído às organizações. Por outro lado, o sufixo é atribuído pela biblioteca digital, sendo que foi projetado para ser flexível com os padrões de identificação da editora (por exemplo, o DOI 10.1000/123456x é composto pelo prefixo 10.1000 e sufixo 123456x). Identificar a publicação *on-line* usando o DOI proporciona acesso eletrônico a todos os arquivos complementares *on-line* associados à publicação.

É importante notar que trabalhos anteriores foram focados na análise do impacto de produções

Figura 2.6: Representação gráfica do problema de predição de ligações.

bibliográficas, enquanto o nosso trabalho é focado na automação da estimativa da IPR.

2.5 Predição de ligações

O problema da predição de ligações é definido matematicamente da seguinte forma:

Definição:. *Seja um conjunto de instâncias $V = \{v_i\}_{i=1}^n$ organizadas em forma de rede social $G = (V, E)$ sendo que E representa o conjunto de vínculos observados. A predição de ligação é a tarefa de prever a probabilidade de um vínculo não observado $e_{i,j}$ existir entre um par arbitrário de vértices i e j com $i, j \in 1..n$ na rede de dados.*

A Figura 2.6 mostra graficamente a tarefa de prever com precisão as ligações (arestas) para uma rede de seis vértices em um tempo t_{n+1} , considerando as ligações da mesma rede em períodos de tempo prévios $t_1, \dots, t_n$. A dificuldade do problema reside no dinamismo típico destas redes cujos nós e arestas mudam constantemente.

Existe uma ampla variedade de propostas no escopo de predição de ligações, sendo que existem também diversas classificações em diferentes taxonomias segundo os levantamentos de informações [Al Hasan e Zaki, 2011; Lü e Zhou, 2011] que serão apresentadas nesta seção.

A primeira corrente de propostas baseia-se no cálculo de algum tipo de medida de similaridade, em que cada par de vértices x e y é atribuído com um valor $sim(x, y)$, o qual, às vezes, é chamado de proximidade entre x e y . Todas as arestas ainda não observadas na rede são ordenadas segundo a medida de similaridade calculada, portanto, arestas com valor de similaridade alto deveriam possuir maior probabilidade de existir. A escolha da medida de similaridade, por sua vez, é outro grande desafio. A este respeito, dependendo do tipo de medida de similaridade usado, estes modelos podem ser subclassificados em algoritmos baseados na similaridade dos atributos próprios dos vértices ou similaridade estrutural, comumente vinculados a características das arestas, que por sua vez, podem ser classificados em algoritmos baseados em similaridade local ou global, dependentes dos vértices ou das arestas, etc. [Al Hasan *et al.*, 2006; Hu *et al.*, 2012; Ströele *et al.*, 2013; Yang *et al.*, 2012].

A Figura 2.7 mostra graficamente a taxonomia dos principais modelos de predição de ligações.

Figura 2.7: Principais modelos propostos para predição de ligações.

Dentre as propostas baseadas em similaridade local mais conhecidas, podemos mencionar o preditor Adamic-Adar [Adamic e Adar, 2003], que faz uma normalização do número de vizinhos em comum de pares de vértices, privilegiando a exclusividade dos vértices. Um outro preditor desse tipo é o baseado na atribuição preferencial [Barabási e Albert, 1999; Hu *et al.*, 2012], o qual propõe outorgar probabilidade de ligação proporcional ao grau dos vértices em questão. Por outro lado, entre medidas de similaridade global, temos o caminho mais curto entre pares de vértices, usado com sucesso também na detecção de comunidades [Yan e Gregory, 2012]. O índice Katz [Katz, 1953] é baseado no cálculo do caminho mais curto entre dois vértices. É realizada a soma direta ponderada de todos os caminhos. Na literatura, esses modelos também são chamados de modelos não supervisionados. Por outra vez, os modelos supervisionados usam modelos de aprendizado para prever ligações usando como base alguns dos preditores não supervisionados como características. Esses métodos otimizam uma função criada com diversos parâmetros, os quais adequam-se melhor aos dados observados. Cada uma das medidas de similaridade são usadas como características para prever ligações entre dois vértices [Al Hasan *et al.*, 2006; Bilgic *et al.*, 2007; Doppa *et al.*, 2009; Lichtenwalter *et al.*, 2010; Potgieter *et al.*, 2009].

Os modelos probabilísticos, por sua vez, visam a abstrair a estrutura da rede observada e fazem a predição de arestas ausentes usando algum modelo de aprendizado, isto é, dada uma rede, o modelo probabilístico tentará otimizar uma função construída para descrever um modelo composto por um conjunto de parâmetros que poderiam ajustar os dados observados da rede a ser analisada. Então, a probabilidade de existência de uma aresta ainda não observada é determinada por uma probabilidade condicional. A essa categoria pertencem os modelos probabilísticos relacionais que são ferramentas de modelagem que fornecem uma maneira sistemática para incorporar tanto atributos dos vértices como das arestas para modelar a distribuição de probabilidade conjunta. A principal vantagem desses modelos em relação a outros modelos - que tentam incorporar ambos tipos de

atributos, de vértices e arestas- é que eles são independentes dos dados, portanto, mais genéricos, e podem ser aplicados em diferentes cenários. Dentre os mais representativos, temos o proposto por Getoor *et al.* [2003], baseado nas redes Bayesianas e que modela as ligações relacionais usando grafos dirigidos acíclicos; [Taskar *et al.*, 2002, 2003] considera essas ligações como não dirigidas e baseia seu modelo nas redes relacionais de Markov. Também devemos mencionar o modelo DA-PER (do inglês *Directed Acyclic Probabilistic Entity Relationship*) proposto por Heckerman *et al.* [2004]. Outro aporte importante destes modelos é que conseguem caracterizar redes heterogêneas [Sun *et al.*, 2012; Yu *et al.*, 2007]. A ideia principal dos modelos probabilísticos bayesianos é a obtenção de uma probabilidade posterior que denota a possibilidade de co-ocorrência de pares de vértices em que estamos interessados. Uma vantagem desse modelo é que o próprio valor pode ser usado como característica em um modelo de classificação. Os modelos básicos usam MRF (do inglês *Markov Random Field*) baseado em probabilidades locais [Kashima e Abe, 2006; Wang *et al.*, 2007] enquanto outros ampliam o modelo para considerar a evolução no tempo, assim como sub-divisões de vértices em grupos para tentar inferir a estrutura hierárquica da rede [Davis *et al.*, 2011; de Sá e Prudêncio, 2011; Popescul e Ungar, 2003; Shin *et al.*, 2012; Sun *et al.*, 2011; Yu *et al.*, 2012].

Um método mais recente é o algébrico linear, que usa fatoração de matrizes e tensores para prever ligações em redes dinâmicas [Acar *et al.*, 2009; Dunlavy *et al.*, 2011; Menon e Elkan, 2011], e também o método baseado na estimação de máxima verosimilhança, o qual cria algumas regras e parâmetros obtidos pela maximização da verosimilhança da estrutura observada e os aplicam na predição de ligações não observadas [Clauset *et al.*, 2008; Guimera *et al.*, 2006; Sun *et al.*, 2011, 2012; Yu *et al.*, 2012].

A predição de ligações e a detecção de comunidades tratam o problema de manipular informações das arestas nas redes. Nesse sentido, as duas tarefas analisam diferentes tipos de atributos [Zhou *et al.*, 2009b]. Os mais utilizados são os atributos estruturais, relacionados à conexidade dos vértices, juntamente com os atributos grupais, os quais contêm contagens vinculadas a grupos aos quais pertencem os vértices. A vantagem de se usar esses tipos de atributo é que eles permitem analisar qualquer tipo de problema que possa ser modelado como grafo. Por outro lado, o uso de atributos descritivos, relacionados a informações próprias dos vértices, são considerados nas análises de problemas específicos. Nos últimos anos, surgiram várias propostas que tentam combinar os diferentes tipos de atributo.

Uma generalização dos preditores não-supervisionados incorporando o peso das arestas é apresentada em [Lü e Zhou, 2009]. Nesse artigo são apresentadas variações no cálculo de alguns dos índices mais conhecidos: número de vizinhos em comum, Adamic-Adar e alocação de recursos. Outra proposta que incorpora atributos apenas nos vértices é apresentada no artigo de Gong *et al.* [2011], no qual os autores incorporam uma estrutura denominada “*Social-Attribute Network*”, a qual cria vértices adicionais na rede para cada um dos atributos descritivos dos vértices da rede original. Uma proposta que usa dados extraídos da Plataforma Lattes é apresentada em [Ströbele

et al., 2013]. Nesse estudo, foram considerados quatro tipos de relações entre pesquisadores, os quais são representados como matrizes de adjacência. Esses autores propõem um método não supervisionado para predizer/sugerir ligações chamadas de *Ströle Metric*.

2.6 Influência na rede de colaboração

Na análise de redes sociais, um dos conteúdos mais importantes é a significância do vértice que implica a influência de um vértice na rede. O cálculo da influência (medida da importância do pesquisador) baseia-se na relação direta ou indireta de cada vértice com respeito a toda a rede [Freeman, 1977; Newman e Park, 2003; Opsahl *et al.*, 2010]. Normalmente, a quantificação da significância baseia-se em medidas globais. Um dos principais fatores é o peso atribuído à ligação e à distância entre todos os pares de vértices. Entre as abordagens mais famosas temos a centralidade de grau (“*degree centrality*”), a centralidade de proximidade (“*closeness centrality*”), a centralidade de intermediação (“*betweenness centrality*”) e suas diferentes variações.

A abordagem da predição de ligações funciona da seguinte maneira: dada uma rede social em um tempo t , procura-se prever as arestas que serão adicionadas à rede durante o intervalo entre t e um tempo futuro t' [Acar *et al.*, 2009; Liben-Nowell e Kleinberg, 2003; Lü e Zhou, 2009; Potgieter *et al.*, 2009]. Há uma série de abordagens neste âmbito, sendo que alguns dos métodos mais simples usam técnicas não supervisionados, que se baseiam no cálculo de algum tipo de semelhança entre os dois vértices (possíveis futuras arestas). Entre os mais conhecidos, temos o Índice Adamic-Adar [Adamic e Adar, 2003], o Índice de Alocação de Recursos [Zhou *et al.*, 2009a] e a Ligação Preferencial [Barabási e Albert, 1999; Hu *et al.*, 2012; Sun *et al.*, 2012].

Vale a pena notar que, em ambas abordagens, a previsão de ligação e a significância têm em comum o cálculo de medidas de similaridade com base em características estruturais da rede. Enquanto a significância normalmente usa medidas globais, alguns modelos de predição de ligações usam apenas medidas locais [Adamic e Adar, 2003; Barabási e Albert, 1999; Hu *et al.*, 2012; Sun *et al.*, 2012; Zhou *et al.*, 2009a]. Este trabalho está focado na análise de redes de colaboração explorando essa semelhança no cálculo das medidas de similaridade.

Propomos uma nova medida de importância dos vértices em uma rede de colaboração que chamamos de *Influência Colaborativa* no processo de predição de ligações. Esta medida usa uma filosofia semelhante à proposta em [Maucher *et al.*, 2011] na reconstrução de uma rede booleana de dados biológicos que requer a identificação de dependências dentro da rede. Na identificação de dependências, os genes são avaliados dependendo da mudança na saída da função booleana ao modificar o estado de um gene e mantendo fixo o estado dos outros genes. Criamos também um preditor de ligações com base em medidas locais e estruturais, que analisa a influência de colaboração de um vértice. A influência é explorada com base na variação da previsão de ligação com a presença ou ausência do vértice em toda a rede. Neste sentido, a influência colaborativa leva em conta a variação do número de ligações adicionadas ou removidas pela previsão, causada pela

remoção do vértice e o desempenho do preditor.

Recentemente, vários trabalhos têm sido dedicados a quantificar e realçar os colaboradores academicamente mais influentes [Abbasi, 2013; Opsahl *et al.*, 2010; Wang e Groth, 2010; Wei *et al.*, 2013; Yan *et al.*, 2013; Zhai *et al.*, 2013]. Para medir isso são utilizados indicadores bibliométricos, como o índice-h e suas variações [Abbasi *et al.*, 2010; Hirsch, 2005] e medidas de análise de redes sociais de centralidade [Freeman, 1977; Opsahl *et al.*, 2010; Wasserman, 1998]. Os resultados obtidos nesses estudos têm demonstrado a aplicabilidade destas medidas em redes de coautoria para refletir o desempenho dos pesquisadores com base em sua posição e influência dentro de sua rede de colaboração. A significância de um vértice em análise de redes sociais é normalmente medida de acordo com os laços que entram ou saem do vértice e a contribuição do vértice no cálculo dos caminhos mais curtos entre todos os pares de vértices. Por causa da natureza de nossa proposta - com base nos vizinhos em comum, seus diferentes tipos de normalizações e o caminho mais curto -, podemos compará-la com algumas medidas de centralidade de vértices bem conhecidas. A medida mais simples e intuitiva de centralidade é a centralidade de grau, que é uma contagem do número de arestas adjacentes a um vértice. Esta medida usa apenas uma generalização desta medida para redes ponderadas, é o *Strength*, que leva em conta o peso das bordas. Com base nessas medidas, Freeman [1977] propôs algumas medidas de centralidade: a centralidade de proximidade e a centralidade de intermediação.

A centralidade de proximidade foi definida como a inversa da *farness* de um vértice u , que é a soma das distâncias para todos os outros vértices da rede. Esta medida foi originalmente proposta para grafos conexos. Se o grafo possui várias componentes conexas, o algoritmo calcula separadamente a medida para cada componente. Algumas variações desse algoritmo que fazem diferentes extensões da medida podem ser encontrados em [Newman, 2001; Opsahl *et al.*, 2010]. A centralidade de intermediação calcula a proporção de caminhos mais curtos entre pares de vértices que passam pelo vértice de interesse. Existem também outras medidas relacionadas ao nosso método. A excentricidade é definida como a distância máxima a partir do vértice de interesse para todos os outros vértices da rede. No caso da medida *communicability centrality* [Estrada e Hatano, 2008; Estrada e Rodriguez-Velazquez, 2005], também chamada de centralidade de subgrafo, um vértice é definido como a soma dos comprimentos de todos os caminhos fechados com início e fim neste vértice. Finalmente, a *Closeness vitality* também foi levada em conta por causa da sua filosofia, que é baseada na remoção de vértices. Esta medida calcula a mudança na soma das distâncias entre todos os pares de vértices, excluindo o vértice em questão.

A seguir, apresentamos a definição formal de algumas das medidas de centralidade mais conhecidas e que serão tomadas como referência para avaliar nossos resultados.

Centralidade de grau

Esta medida é a mais intuitiva e simples. A centralidade de um vértice é igual ao número de contatos diretos que ele possui; ela foi inicialmente proposta por [Shaw, 1964] e depois usada como referência para propor novas medidas de similaridade por [Freeman, 1979]. Mais formalmente, seja $G = (V, E)$ um grafo, e seja v_k um vértice de G ($v_k \in V$). A centralidade de grau de V_k , denotada por d_k , é o número de arestas incidentes a v_k . Isto é,

$$d_k = \sum_{j=1}^n e_{kj}, \quad (2.6)$$

em que e_{kj} são arestas adjacentes a v_k .

Centralidade de proximidade

Diferentemente da medida anterior, esta medida de centralidade considera a existência de alguma métrica entre pares de vértices. Na proposta inicial, o comprimento de seus caminhos mais curtos é usado como métrica [Sabidussi, 1966]; é importante notar que a soma dos caminhos mais curtos de um vértice para todos os outros vértices do grafo é chamado de afastamento. Portanto, a proximidade é definida como o inverso do afastamento. Então, seja um grafo $G = (V, E)$ conexo com n vértices, e seja v_k um vértice de grafo ($v_k \in V$). A centralidade de proximidade de v_k está definida pelo inverso de soma das distâncias de v_k a todos os demais vértices do grafo,

$$c(v_k) = \frac{1}{\sum_{j=1}^n \text{dist}(v_j, v_k)}, \quad (2.7)$$

Centralidade de intermediação

Um dos grandes desafios das redes sociais é que elas são muito grandes e, em muitos casos, desconexas. As medidas de centralidade existentes até a década dos 70 eram limitadas a redes conexas; então, [Freeman, 1977] propôs uma medida baseada na influência que um vértice poderia exercer sobre os outros vértices em uma rede real. Formalmente, seja um grafo $G = (V, E)$ com n vértices, e seja v_k um vértice de G , a centralidade por intermediação está definida por:

$$c(v_k) = \sum_{v_i \neq v_k \neq v_j} \frac{\sigma_{v_i, v_j}(v_k)}{\sigma_{v_i, v_j}}, \quad (2.8)$$

em que σ_{v_i, v_j} é o número total de caminhos mais curtos entre os vértices v_i e v_j e $\sigma_{v_i, v_j}(v_k)$ é o número desses caminhos que passam pelo vértice v_k .

Centralidade de autovetor

A medida foi proposta por [Bonacich, 1987], baseada em autovalores e autovetores de matrizes simétricas, cujo objetivo é medir a importância de um vértice em função da importância de seus vizinhos. Matematicamente, seja um grafo conexo $G = (V, E)$ com n vértices e seja v_k um vértice de G , a centralidade de vetor próprio de v_k está definida como:

$$c(v_k) = x_k, \quad (2.9)$$

em que k_x é a k -ésima coordenada do vetor próprio positivo unitário x associado ao índice do grafo, isto é,

$$x_k = \frac{1}{\rho} \sum_{j=1}^n a_{kj} x_j; k = 1, \dots, n, \quad (2.10)$$

em que a_{kj} são elementos da matriz de adjacência $A(G)$.

Capítulo 3

Modelagem de redes usando ARGs

Neste capítulo é detalhada a descrição da abordagem proposta, levando em conta as definições apresentadas no Capítulo 2 e considerando a disponibilidade pública de informações dos pesquisadores brasileiros na plataforma Lattes.

3.1 Método proposto

Neste trabalho, motivado pela interação através de trabalhos colaborativos e a assortatividade entre atributos dos pesquisadores, é proposto o uso simultâneo de atributos estruturais e atributos descritivos através de grafos relacionais com atributos (ver Figura 3.1), com o intuito de melhorar o estudo da taxa de internacionalização, o desempenho da predição de ligações e, consequentemente, a análise da influência dos pesquisadores em sua rede de colaboração.

Figura 3.1: Grafo relacional com atributos (ARG)

A Figura 3.2 mostra um esquema de alto nível de nossa proposta, a qual está dividida em três etapas, descritas a seguir:

1. Extração de dados acadêmicos

Figura 3.2: Esquema de alto nível da proposta

Esta primeira etapa está focada na extração automatizada de dados públicos dos pesquisadores a partir da Plataforma Lattes. Para se atingir esse objetivo, será usada a ferramenta *scriptLattes* desenvolvida por Mena-Chalco e Cesar-Jr [2009]. Com a finalidade de extrair os dados necessários para nosso estudo, foi realizada uma pequena modificação no código da ferramenta a fim de extrair, independentemente, os diferentes tipos de colaborações entre pesquisadores.

O *scriptLattes* recebe como entrada a lista dos pesquisadores cujos currículos Lattes serão analisados e o período de tempo que será examinado, para finalmente retornar como saída diversos relatórios acadêmicos. Neste estudo serão avaliados os grafos de colaboração retornados pelo programa, representados por matrizes de adjacência contendo o número de

#vertices						
#vertex_id	#vertex_attribute_label_1	#vertex_attribute_label_2	...#vertex_attribute_i			
01	,	3	,	3.5	,	75
02	,	5	,	3.3	,	10
03	,	6	,	3.9	,	0
04	,	3	,	2.5	,	11
.						
.						
.						
#edges						
#edge_source	#edge_target	#edge_attribute_label_1	#edge_attribute_label_2	...#edge_attribute_label_j		
01	,	02	,	2	,	10
01	,	04	,	4	,	0
03	,	04	,	20	,	6
.						
.						
.						

Figura 3.3: Arquivo padrão para armazenar ARGs

colaborações de cada tipo. O programa foi ligeiramente modificado para gerar um arquivo .ARG com o formato mostrado na Figura 3.3 com i atributos de vértice e j atributos de aresta.

2. Geração de grafos relacionais com atributos

Como descrito no Capítulo 2, os grafos relacionais com atributos fornecem a capacidade de armazenar detalhes adicionais tanto dos pesquisadores (vértices dos grafos) quanto das relações de colaboração entre eles, em vetores de atributos μ e ν respectivamente.

Finalizada a etapa de extração de dados, descrita na seção acima, serão gerados arquivos .ARG contendo todas as informações das arestas e vértices. As publicações foram divididas em sete categorias e serão carregadas no vetor ν , enquanto os atributos individuais serão carregados no vetor μ , como mostrado na Figura 3.4.

Todos esses dados também serão obtidos usando o scriptLattes. Especial atenção merece a taxa de internacionalização cujo método de cálculo será apresentado na Seção 3.2 deste Capítulo.

Podemos considerar como uma subetapa o procedimento pelo qual colapsamos vários grafos em um único grafo resumo correspondente a um período de vários períodos, como é mostrado na Figura 3.5.

3. Análise de redes de colaboração científica

O objetivo deste trabalho é proporcionar um modelo que caracterize as redes de colaboração científica de maneira realista, portanto, deve permitir gerar soluções eficientes a diversas

Figura 3.4: ARG mostrando os atributos das arestas (tipo de publicações) e dos vértices.

Figura 3.5: Processo para colapsar várias ARGs em um único ARG resumo.

aplicações no escopo da análise de redes de colaboração científica. Mantendo todos os dados em uma rede única, será possível fazer diversos tipos de análises, em nível individual ou coletivo, isto porque será armazenada uma grande quantidade de dados e a estrutura é muito flexível, tornando a proposta genérica. Com a finalidade de avaliar o modelo proposto, foram escolhidas três aplicações de alto interesse atualmente:

Taxa de internacionalização: na Seção 3.2 será apresentado um método para calcular, de maneira automatizada, a taxa de publicações internacionais. Este valor é computado considerando apenas um tipo de publicação (em revistas científicas) e para um grupo de pesquisadores. Com o propósito de usar essa taxa como um atributo descritivo de cada um, o método de cálculo deverá ser modificado de forma que o valor de internacionalização seja calculado para cada um e considerando os diferentes tipos de publicações. Isto permitirá obter resultados mais exatos de internacionalização, além de permitir fazer análises a nível individual ou coletivo, assim como pelos tipos de publicações.

Predição de ligações: essa aplicação é análoga a que foi mostrada nos trabalhos de Gong

et al. [2011]; Yang *et al.* [2011]; Zhou *et al.* [2009b] nos quais foram incluídos com sucesso atributos sociais ou descritivos na análise de redes sociais. Neste projeto, pretende-se incluir os três tipos de atributos - descritivos, estruturais e grupais - com a finalidade de melhorar o desempenho do método de predição de ligações. Foi escolhido um método de aprendizado supervisionado, mas dada a flexibilidade dos ARGS, é possível aplicar outros algoritmos de predição para avaliar o impacto da inclusão dos novos atributos no resultado da predição. Este estudo será dinâmico, sendo que a antiguidade será considerada como uma variável.

Com a inclusão dos novos atributos, será possível estabelecer a influência da taxa de internacionalização na predição de ligações. Também poderemos determinar a influência de cada um dos tipos de relações na predição global, além de determinar o tipo específico da ligação prevista.

Influência na rede de colaboração: esta proposta está baseada no resultado da predição de ligações e representa uma das principais contribuições deste trabalho. Será analisada a influência de cada um dos pesquisadores, na rede contemplando a mudança nos resultados causada pela presença ou ausência do pesquisador (vértice) na rede, durante o processo de predição. Esta é uma nova proposta que permitirá medir quantitativamente a influência dos pesquisadores na rede. Finalmente, com a finalidade de diminuir o custo computacional no processo de treinamento do preditor são usadas principalmente medidas locais e o algoritmo de treinamento é adaptado.

A seguir são apresentados os métodos propostos para as três aplicações específicas que permitirão realizar os experimentos para avaliar o modelo proposto. Também são apresentados os resultados experimentais obtidos para cada um dos métodos propostos.

3.2 Análise da taxa de internacionalização

Nos últimos anos, o Brasil tem registrado um crescimento constante no domínio da investigação científica, trazendo, como consequência natural, um aumento em publicações científicas, além do fortalecimento das relações com os países tradicionalmente colaboradores e da criação de novas relações com outros países. Isso trouxe uma clara tendência para o trabalho colaborativo com pesquisadores estrangeiros, criando a necessidade de acompanhamento, avaliação e análise dessa tendência. Há estudos anteriores voltados para realização desta tarefa, como [Glänzel *et al.*, 2006; Leta *et al.*, 2006], que fazem a análise da evolução da atividade publicação e do impacto de citações em instituições brasileiras de pesquisa com base nos dados bibliográficos ISI.

Em [Leite *et al.*, 2011], a medida IPR, a qual é usada em nosso trabalho, foi proposta, e um subconjunto do banco de dados do CNPq (Lattes) foi usado como fonte de dados. Os autores

Figura 3.6: passos seguidos para o cálculo do IPR, em termos de coautoria internacional, a partir de produções bibliográficas publicadas por um grupo de pesquisa.

fizeram um estudo detalhado das influências de parâmetros distintos tais como os setores, área da carreira, idade e gênero nas publicações internacionais de pesquisadores brasileiros.

Neste trabalho, a avaliação quantitativa da IPR implicou quatro passos: (i) a geração de produções bibliográficas únicas; (ii) a seleção de produções bibliográficas em que o DOI foi informado pelos pesquisadores; (iii) a extração de endereços dos coautores; e (iv) a identificação dos nomes dos países dos colaboradores.

A Figura 3.6 representa graficamente os passos do nosso método proposto para cálculo da IPR, em termos de coautoria internacional a partir de produções bibliográficas publicadas por um grupo de pesquisa.

Em nosso trabalho, as publicações de autores foram extraídas de seus Currículos Lattes usando scriptLattes [Mena-Chalco e Cesar-Jr, 2009] que gera uma lista de publicações. Normalmente, é comum que nos grupos de pesquisa algumas das publicações sejam elaboradas em coautoria com diferentes pesquisadores do mesmo grupo. Assim, a mesma publicação acadêmica pode ser registrada em todos os currículos Lattes dos coautores envolvidos. O processo de tratamento de redundância permite a detecção de publicações iguais. Neste contexto, as publicações são usadas para detectar a coautoria entre os membros do grupo de pesquisa: dois ou mais membros são considerados coautores, se houver uma produção comum (ou igual) entre eles [Mena-Chalco e Cesar-Jr., 2011].

É interessante notar que, nem todas as publicações têm registrado o DOI na plataforma Lattes, pois o DOI é manualmente informado pelo pesquisador. Assim, o processo de preenchimento é suscetível a erros, de modo que o procedimento desenvolvido verifica e seleciona exclusivamente

publicações nas quais o DOI é informado pelos pesquisadores. Para obter as informações de endereço dos coautores, os DOIs de suas publicações foram usados da seguinte maneira: para cada publicação da lista (compilada conforme indicado acima) está ligado um DOI único, que é usado como um *link* (URL) para obter a informação eletrônica da publicação. Toda biblioteca digital tem atribuído um prefixo único do DOI (por exemplo, o prefixo 10.1016 é de propriedade da Science Direct www.sciencedirect.com) e mantém o seu próprio modelo de página *web* para mostrar as informações mais importantes das produções bibliográficas publicadas (essa informação é apresentada em um formato estruturado ou semi-estruturado). Então, o nosso principal desafio foi o de examinar as páginas *web* a fim de extrair, automaticamente, as informações dos endereços dos coautores das bibliotecas digitais.

Foi implementado um analisador de página *web* personalizado (*parser*) [Tomita, 1991], o qual permite facilmente configurar e adicionar um novo analisador de página *web*. Esse analisador extrai informações sobre o endereço dos autores, procurando o *Tag HTML* específico, indicado no analisador. A escolha do analisador apropriado para cada publicação depende do prefixo DOI.

Além disso, um dicionário de 236 nomes de países foi criado devido ao fato de que a maioria dos países apresenta nomes diferentes, em diferentes idiomas, bem como, diferentes representações (por exemplo, “Brasil” está relacionado ao “Brazil”, já “USA”, “América”, “US”, “U.S.” e “EUA” estão relacionados aos “Estados Unidos”).

Na primeira etapa, os nomes dos países padrões (versão em inglês) são pesquisados no endereço dos coautores selecionados, anteriormente normalizados para manter os mesmos caracteres de codificação.

A identificação dos nomes dos países é realizada usando 15 expressões regulares (representada no analisador), cada uma projetada para o processamento de página *web* de uma biblioteca digital específica, ou seja: ACM Digital Library, IEEE Xplore Digital Library, Science Direct, Springer-Link, Proceedings of the National Academy of Sciences, Online Library Wiley, Journal of Applied Physics, Optics Info Base, Physical Review, IOPscience, BioMed Central, Molecules, The Journal of Biological Chemistry, ACS Publications e Scielo. Essas bibliotecas foram escolhidas do *ranking* das mais usadas para publicar as produções dos pesquisadores brasileiros. Além disso, uma expressão regular genérica foi utilizada nos casos em que o DOI não estivesse relacionado a qualquer uma das bibliotecas acima referidas. Um *parser* genérico foi desenvolvido para procurar nomes de países sem restrições de Tags HTML específicas. Com este procedimento, é possível associar um ou mais nomes de países dos coautores que participaram das publicações acadêmicas.

Finalmente, note que existem alguns casos em que o DOI não poderia ser examinado devido a diferentes razões, tais como: a URL está quebrada ou o número DOI foi mal escrito, ou a página *web* das bibliotecas não contém o endereço dos autores. Esta questão representa um desafio para identificar corretamente os nomes dos países dos coautores envolvidos na lista de publicações.

O algoritmo a seguir resume o procedimento utilizado para identificar a identidade das colaborações internacionais a partir de uma lista de produções bibliográficas:

IDENTIFY-INTERNATIONAL-PUBLICATIONS(*Group*)

```

1 PublicationList ← GET-PUBLICATION-LIST(Group)
2 for each publication ∈ PublicationList
3   do
4     if HAS-REGISTERED-DOI(publication)
5       then ▷ Just publications with registered DOI
6         for each country ∈ CountryDictionary
7           do
8             if SEARCH-COUNTRY(publication, country)
9               then
10                 ADD(InternationalPublicationList
11                   [publication], country)
11 return InternationalPublicationList

```

O método IDENTIFY-INTERNATIONAL-PUBLICATIONS requer a informação do grupo de pesquisa a fim de identificar as publicações internacionais. O procedimento GET-PUBLICATION-LIST permite obter as produções bibliográficas publicadas pelo grupo. O método HAS-REGISTERED-DOI verifica se a publicação atual tem o DOI. O método SEARCH-COUNTRY utiliza o DOI para acessar o conteúdo eletrônico das publicação, bem como para selecionar o *parser* a fim de extrair os endereços de autores da página *web*.

3.2.1 Descrição dos dados

Neste trabalho, utilizou-se a produção bibliográfica registrada na Plataforma Lattes como fonte de dados. O conjunto de dados foi extraído da Plataforma Lattes em maio de 2011 e contém cerca de 1,3 milhão de currículos.

Pesquisadores de cinco grandes áreas de conhecimento foram consideradas no presente estudo: (i) Ciências Agrárias, (ii) Ciências Biológicas, (iii) Ciências Exatas e da Terra, (iv) Ciências da Saúde e (v) Engenharia. A fim de padronizar nossa análise, apenas os artigos publicados em revistas científicas entre 1990 e 2010 foram considerados. A Tabela 3.1 apresenta a quantidade de pesquisadores para as diferentes áreas de conhecimento. A classificação das grandes áreas do conhecimento foi adotada a partir da agência de financiamento de pesquisa CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). Note que, além de uma área do conhecimento, a IPR também pode ser atribuída a cada pesquisador. O conjunto de dados considerado contém registros de produções bibliográficas (em revistas científicas) extraídas da plataforma Lattes para o período entre 1990 e 2010.

3.2.2 Resultados

Um total de 790 921 currículos Lattes foram analisados para finalmente se selecionar, para cada área do conhecimento, mais de 100 pesquisadores de colaboração, ou seja, foram selecionados os 100 pesquisadores mais trabalhos colaborativos.

Área de conhecimento	Pesquisadores
Ciências Agrárias	92 927
Ciências Biológicas	128 104
Engenharia	120 993
Ciências Exatas e da Terra	176 114
Ciências da Saúde	272 783

Tabela 3.1: *Quantidade de pesquisadores para as áreas de conhecimento consideradas.*

Área	P	P-DOI	P-I (IPR)
Ciências Agrárias	13 784	2 894 (20,99%)	220 (1,59%)
Ciências Biológicas	12 977	4 557 (35,11%)	1 013 (7,80%)
Engenharia	9 349	3 668 (39,23%)	467 (4,99%)
Ciências Exatas e da Terra	13 846	5 202 (37,57%)	1 067 (7,70%)
Ciências da Saúde	19 376	4 257 (21,97%)	668 (3,44%)

Tabela 3.2: *Resumo dos resultados obtidos para os grupos brasileiros de pesquisa, associados com cinco grandes áreas do conhecimento. P indica número de publicações, P-DOI indica número de publicações com DOI registrado, P-I indica número de publicações internacionais. O valor entre parênteses indica a proporção em relações a P-DOI.*

Um resumo dos resultados obtidos para os grupos brasileiros de pesquisa, associados com as cinco áreas do conhecimento no período de avaliação, é apresentado na Tabela 3.2. Esta tabela mostra que o total de publicações (P), em todos os casos, é significativamente maior do que as publicações com DOI registrado (P-DOI). Observa-se que a percentagem de publicações na qual o número DOI foi registrado é baixa e corresponde ao intervalo de 20,99 % (Ciências Agrárias) a 39,23 % (Engenharia).

Na tabela, as publicações internacionais (P-I) são mostradas na terceira coluna. Note-se que, considerando os mais de 100 pesquisadores de colaboração, os grupos com maior IPR pertencem às Ciências Biológicas (7,80 %) e às Ciências Exatas e da Terra (7,70 %). Por outro lado, o grupo de Ciências Agrárias apresenta a IPR mais baixa (1,59 %).

Os resultados apresentados aqui reforçam a evidência, discutida na literatura [Leta *et al.*, 2006], que indica que os pesquisadores brasileiros preferem publicar com pesquisadores brasileiros (colaboração nacional). Este cenário é persistente ao longo do tempo, tal como demonstrado na análise seguinte.

A fim de investigar a percentagem de publicações em revistas internacionais em diferentes períodos de tempo, as informações dos artigos são discretizadas por ano.

Figura 3.7: Medidas quantitativas acadêmicas obtidas com os 100 pesquisadores mais colaborativos para cinco áreas de conhecimento: Ciências Agrárias, Ciências Biológicas, Engenharia, Ciências Exatas e da Terra e Ciências da Saúde. Apenas artigos publicados em revistas entre 1990 e 2010 foram considerados.

A Figura 3.7a mostra a quantidade de artigos internacionais publicados em revista, entre 1990 e 2010, pelos pesquisadores mais colaborativos nas principais áreas do conhecimento. Observou-se que cada grupo de pesquisa apresenta um comportamento semelhante no crescimento das publicações ao longo dos anos. Além disso, observa-se que o número de artigos de revistas tem experimentado um padrão de crescimento linear. É interessante notar ainda que as pesquisas consideradas associadas às Ciências da Saúde mostram um crescimento acentuado, se comparado a outras áreas. Por outro lado, os pesquisadores associados à área da Engenharia mostram o menor crescimento.

A Figura 3.7b apresenta a quantidade de artigos publicados em revistas científicas no mesmo período de tempo em que o DOI foi registrado na Plataforma Lattes por grupos de pesquisa. Note que, para cada área, a quantidade de publicações cresce seguindo a função logística de crescimento, em que a fase inicial é um crescimento exponencial, em seguida, o crescimento diminui. Conforme relatado anteriormente, em 2010, as pesquisas associadas às Ciências da Saúde apresentaram a maior quantidade de publicações nas quais os DOI eram informados.

A percentagem de crescimento de publicações internacionais para cada grupo é apresentada na Figura 3.7(c). Como pode ser visto, em todos os casos, a percentagem de publicações internacionais aumenta ao longo do tempo. Também é possível ver um platô entre 1993 e 2001, como descrito em [Leta e Chaimovich, 2002]. Os grupos de pesquisa associados às Ciências Biológicas e às Ciências Exatas e da Terra apresentam um comportamento semelhante no crescimento do percentual internacional de publicações ao longo do tempo. Embora o número total de publicações do Brasil seja cada vez maior, as publicações internacionais têm crescido mais rápido. Observe que, em todos os casos, depois de 2001, a percentagem de publicações internacionais tem um crescimento exponencial.

De fato, os grupos associados às Ciências Biológicas e às áreas de Ciências Exatas e da Terra apresentam maior percentagem de internacionalização, enquanto o grupo associado à área de Ciências Agrárias apresenta menor percentual (IPR).

O número de países colaboradores com os quais trabalham os coautores brasileiros é apresentado na Figura 3.7d. O número de diferentes países, presentes na lista de coautorias, tem aumentado ao longo do tempo. De fato, para cada grupo de pesquisa, o número de países estrangeiros segue um padrão de crescimento linear. Além disso, é interessante notar que o número de países estrangeiros mostra um crescimento proporcional se comparado com os valores de IPR computado para todo o período de avaliação (ver Tabela 3.2). Além disso, nos últimos anos, os grupos associados às Ciências Biológicas, às Ciências da Saúde e às Ciências Exatas e da Terra se tornaram mais internacionais.

A tabela 3.3 apresenta uma lista com os *top 30* dos países que mais colaboram com pesquisadores brasileiros em publicações em revistas. Por exemplo, o país mais frequentemente identificado nas publicações internacionais são os Estados Unidos para os grupos de pesquisa associados às Ciências Agrárias, às Ciências Biológicas, às Ciências Exatas e da Terra e às Ciências da Saúde.

É interessante notar que as publicações entre dois países estrangeiros diferentes, não incluindo o Brasil, não foram considerados neste estudo. Países como os Estados Unidos e o Canadá, na América do Norte, a França, o Reino Unido; e a Espanha, na Europa, e a Argentina na América Latina são os colaboradores mais frequentes com os pesquisadores brasileiros nas cinco áreas analisadas neste trabalho. A Figura 3.8 apresenta uma representação gráfica das colaborações em todo o mundo.

Ranking	Ciências Agrárias		Ciências Biológicas		Engenharia		Ciências Exatas e da Terra		Ciências da Saúde	
	País	Nº	País	Nº	País	Nº	País	Nº	País	Nº
1	USA	83	USA	343	Spain	93	USA	168	USA	260
2	Germany	26	UK	172	France	70	Spain	137	UK	112
3	Argentina	21	France	145	Germany	69	France	123	Canada	72
4	UK	19	Germany	100	Portugal	65	Argentina	115	Italy	58
5	Spain	16	Italy	75	USA	55	Germany	109	France	57
6	Canada	14	Canada	42	Argentina	37	Japan	75	Spain	44
7	Denmark	13	Australia	40	UK	25	UK	72	Germany	43
8	France	12	Argentina	37	Italy	19	India	58	Australia	30
9	Italy	9	Spain	32	Canada	19	Portugal	57	Argentina	28
10	Australia	7	Switzerland	32	Colombia	17	Canada	43	Switzerland	26
11	Portugal	6	Japan	29	India	17	Italy	41	Portugal	23
12	Egypt	4	Portugal	27	Serbia	13	Uruguay	35	Japan	22
13	Uruguay	4	Mexico	25	Sweden	12	Belgium	29	Mexico	19
14	Chile	3	Uruguay	22	Uruguay	11	Chile	29	Chile	17
15	Colombia	3	Denmark	20	Denmark	10	Australia	26	New Zealand	16
16	Cuba	3	Chile	18	Japan	10	Cuba	25	Sweden	16
17	Costa Rica	2	Colombia	17	Cuba	8	Mexico	20	Colombia	11
18	Finland	2	Belgium	15	Mexico	7	Poland	18	Poland	9
19	Guadeloupe	2	Nigeria	14	Russia	7	Russia	15	Finland	8
20	Japan	2	Venezuela	13	Bulgaria	6	Sweden	14	South Africa	8
21	Nigeria	2	Sweden	11	Belgium	5	Serbia	13	Belgium	7
22	Switzerland	2	India	10	Australia	4	Colombia	11	Israel	7
23	Ecuador	1	Singapore	9	Georgia	4	Denmark	9	Netherlands	7
24	El Salvador	1	Austria	8	Switzerland	4	Finland	8	Peru	7
25	Kenya	1	Pakistan	8	Chile	3	Malaysia	8	Taiwan	7
26	Netherlands	1	South Africa	8	China	3	Puerto Rico	7	Turkey	7
27	Pakistan	1	China	7	Peru	3	China	6	Ecuador	6
28	Paraguay	1	Norway	7	Montenegro	2	Peru	6	Russia	6
29	Peru	1	Peru	7	Slovenia	2	Romania	6	Norway	6
30	South Africa	1	Ireland	6	Ukraine	2	Ukraine	5	Thailand	4

Tabela 3.3: O top dos 30 países que mais colaboram com o Brasil. A quantidade numérica indica o número de publicações.

Figura 3.8: Mapas mundiais de colaborações acadêmicas para as diferentes áreas (Ver Tabela 3.3).

3.3 Predição de ligações

É possível modelar o problema de predição de relações como uma tarefa de classificação supervisionada, no qual cada ponto no espaço de características corresponde a um par de vértices na rede científica.

Considere a rede científica $G = (V, E)$, na qual os vértices $v_i \in V$ representam os pesquisadores, e as arestas $(v_i, v_j) \in E$ representam a relação de colaboração entre os pesquisadores denotados por v_i e v_j em um período particular de tempo t . Para cada par de vértices, são calculados os atributos que poderiam estar correlacionados com a probabilidade de existir uma relação entre dois vértices no futuro. O conjunto de atributos calculados para um par de vértices formará o vetor de características. f_{ij} denota o vetor de características para o par de vértices v_i e v_j . A Figura 3.9 mostra graficamente este processo de construção do preditor de ligações a partir de uma rede de colaboração científica.

Figura 3.9: Representação gráfica do procedimento de construção do preditor de ligações.

O conjunto de publicações foi dividido em intervalos de tempo, não sobrepostos, relacionados com o ano de publicação, tendo sido criadas redes de colaboração anuais. Essas redes foram agrupadas em su-períodos de tempo. Para isso, várias redes anuais foram sintetizadas em uma única rede correspondente ao sub-período em questão. O grafo G'_1 do primeiro intervalo $([t_0, t'_0])$ é usado como conjunto de treinamento para prever as ligações do grafo G'_2 . O segundo grafo, correspondente ao intervalo $([t_1, t'_1])$, é usado como conjunto de teste para prever as ligações do grafo G'_3 , correspondente ao intervalo $([t_2, t'_2])$, como é mostrado na Figura 3.5. A Figura 3.12 mostra graficamente o modelo proposto para prever ligações.

Figura 3.10: Representação gráfica da abordagem de predição de ligações.

Mais formalmente, suponha-se que v_i e $v_j \in V$ sejam dois vértices da rede científica $G(V, E)$ e

o rótulo do ponto no espaço (v_i, v_j) , que representa a aresta entre os vértices v_i e v_j é $\varphi_{(v_i, v_j)}$. Então, este trabalho tem por objetivo criar um preditor que determinará a etiqueta de uma aresta segundo a seguinte função:

$$\varphi_{(v_i, v_j)} = \begin{cases} +1, & \text{se } (v_i, v_j) \in E \\ -1, & \text{se } (v_i, v_j) \notin E \end{cases} \quad (3.1)$$

Note-se que, neste trabalho, a relação entre v_i e v_j é simétrica, então, o par (v_i, v_j) e (v_j, v_i) representam o mesmo ponto, consequentemente, $\varphi_{(v_i, v_j)} = \varphi_{(v_j, v_i)}$.

Assim, o modelo de classificação do problema de predição de relações necessita predizer a relação pelo sucesso na distinção da classe positiva do conjunto de dados. Isto é, predizer o rótulo desconhecido para um par de vértices (v_i, v_j) , tal que $(v_i, v_j) \notin E$, na rede científica para o intervalo $([t_1, t'_1])$.

Vetor de características

A escolha das características adequadas é crítica para qualquer algoritmo de aprendizagem. Existem muitas medidas estruturais que podem ser usadas na construção do vetor de características. A vantagem de usar esse tipo de medida é que elas são genéricas e, portanto, podem ser aplicadas em redes de qualquer domínio. Tipicamente, computam uma medida de similaridade baseada na vizinhança dos vértices ou no agrupamento de caminhos entre pares de vértices. Formalmente, o conjunto de vetores de características é denotado por $F = \{f_{v_i v_j} | (v_i, v_j) \in V \times V\}$.

Neste trabalho, são consideradas redes de colaboração não dirigidas tanto ponderadas como não ponderadas. Foi construída uma rede para cada ano do período de tempo considerado, cada uma representando a interação colaborativa dos pesquisadores no ano em questão. Note que o número de vértices é fixo em todas as redes, por outro lado, o conjunto de arestas varia no tempo, isto é, uma aresta pode existir em um ano, sumir no ano seguinte e reaparecer no próximo.

A escolha das características foi baseada nos resultados experimentais, levando-se em conta que, neste trabalho, tenta-se usar prioritariamente medidas locais, por causa baixo custo computacional.

Dados dois vértices $v_i, v_j \in V$, $\Gamma(v_i)$ denota o conjunto de vizinhos do vértice v_i , $w(v_i, v_j)$ é o peso da aresta (v_i, v_j) e $s(v_i) = \sum_{r \in \Gamma(v_i)} w(v_i, v_r)$ representa a soma dos pesos das arestas adjacentes ao vértice v_i .

Distância entre dois vértices

A distância entre dois vértices v_i e v_j é definida como o número mínimo de arestas os conectando. Se não existe um caminho que os conecte, então, a distância é assumida como infinita. Mais formalmente:

$$d(v_i, v_j) = \min\{P \mid P = (v_1, v_2, \dots, v_n)\}$$

sendo que $v_i = v_1$ e $v_j = v_n$ e para toda n minimiza:

$$\sum_{i=1}^{n-1} \omega(v_i, v_{i+1})$$

sendo que $\omega(v_i, v_j)$ representa o peso da aresta que liga v_i com v_j .

Existem algoritmos bem conhecidos para calcular esta medida, tendo sido usado o algoritmo Dijkstra. Para as redes ponderadas no cálculo da distância, foi usado o valor inverso do peso original das aresta ($1/w$) para se obter o resultado desejado.

Vizinhos em comum

Em redes não ponderadas, o número de vizinhos em comum para dois vértices v_i e v_j é calculado da seguinte forma:

$$CN(v_i, v_j) = |\Gamma(v_i) \cap \Gamma(v_j)| \quad (3.2)$$

Por outro lado, para redes ponderadas foi usado:

$$CN(v_i, v_j) = \sum_{v_z \in \Gamma(v_i) \cap \Gamma(v_j)} w(v_i, v_z) + w(v_z, v_j) \quad (3.3)$$

A razão para se usar essa medida é que ela evidencia a propriedade de transitividade na rede. Então enquanto o número de vizinhos em comum cresce entre v_i e v_j , a probabilidade de ter uma aresta entre eles no futuro é maior.

Coefficiente de Jaccard

Esta medida é uma normalização do número de vizinhos em comum. Para redes não ponderadas está definida por:

$$JC(v_i, v_j) = \frac{|\Gamma(v_i) \cap \Gamma(v_j)|}{|\Gamma(v_i) \cup \Gamma(v_j)|} \quad (3.4)$$

O Coeficiente de Jaccard define a probabilidade de que um vizinho em comum de v_i e v_j seja selecionado se a escolha é feita aleatoriamente da união dos conjuntos de vizinhos de v_i e v_j .

Portanto, quanto maior o número de vizinhos em comum, maior será o coeficiente de Jaccard.

No caso das redes ponderadas, foi usada a seguinte fórmula:

$$JC(v_i, v_j) = \frac{\sum_{v_z \in \Gamma(v_i) \cap \Gamma(v_j)} w(v_i, v_j) + w(v_z, v_j)}{s(v_i) + s(v_j)} \quad (3.5)$$

Índice Adamic/Adar

Esta medida atribui maior peso aos vértices vizinhos que não são compartilhados com outros vértices [Adamic e Adar, 2003]:

$$AA(v_i, v_j) = \sum_{v_k \in \Gamma(v_i) \cap \Gamma(v_j)} \frac{1}{\log |\Gamma(v_k)|} \quad (3.6)$$

A ideia é que dois vértices que têm um vizinho em comum que não é vizinho de outros vértices são mais semelhantes do que outros dois vértices que têm um vizinho em comum, que também é vizinho de um terceiro vértice.

Para redes ponderadas foi usada:

$$AA(v_i, v_j) = \sum_{v_z \in \Gamma(v_i) \cap \Gamma(v_j)} \frac{w(v_i, v_z) + w(v_z, v_j)}{\log(1 + s(v_z))} \quad (3.7)$$

Ligação preferencial

Esta medida se baseia no modelo “*rich get richer*”, o qual assume que as novas ligações têm maior probabilidade de conectar vértices de graus altos do que vértices de baixo graus. No caso de redes científicas isto significa que é mais provável um novo trabalho colaborativo existir entre pesquisadores que colaboram frequentemente. Assim, para redes não ponderadas foi usada:

$$LP(v_i, v_j) = |\Gamma(v_i)| \cdot |\Gamma(v_j)| \quad (3.8)$$

Para redes ponderadas foi usada:

$$LP(v_i, v_j) = s(v_i) \cdot s(v_j) \quad (3.9)$$

Número de publicações em periódicos

É uma contagem simples do número de publicações em periódicos de cada um dos vértices vinculados em uma relação. Esses atributos são exclusivamente vinculados aos vértices e são diretamente extraídos dos vetores dos vértices dos ARGS.

Número de publicações em congressos

É uma contagem simples do número de publicações em congressos de cada um dos vértices vinculados em uma relação. Esses atributos são exclusivamente vinculados aos vértices e também são diretamente extraídos dos vetores dos vértices dos ARGs.

Taxa de internacionalização

O cálculo dessa medida foi descrito na Seção 3.2, tendo sido calculado para cada vértice e armazenado no correspondente vetor de atributo.

Soma de todas as publicações

Também é uma soma simples do número total de publicações de cada um dos pesquisadores extraída do vetor de atributos dos vértices.

3.3.1 Modelo de classificação

Existe uma grande variedade de modelos de classificação supervisionada que podem ser usados para resolver o problema da predição de ligações: árvores de decisão, naive Bayes, redes neurais, SVM, k vizinhos mais próximos, etc. Para este primeiro protótipo foi escolhido o *Support Vector Machine* SVM [Burges, 1998; Vapnik, 1999], com base nos experimentos realizados.

A classificação supervisionada inicia com um conjunto inicial de dados de treinamento, consistindo de m amostras rotuladas, definidas como:

$$E_m = \{(x_i, y_i) \in R^n \times \{-1, 1\} : i = 1, \dots, m\}$$

Cada x_i é um vetor de características na dimensão d , o qual descreve um ponto no espaço $y_i \in \{-1, +1\}$, e indica se o ponto pertence à classe (+1) ou à classe (-1) que o classificador aprenderá a reconhecer.

O classificador SVM procura um hiperplano nas d dimensões, o qual é uma superfície $(d - 1)$ -dimensional definida como:

$$f(x) = \mathbf{w} \cdot x + b = 0 \quad (3.10)$$

sendo que $\mathbf{w}$ denota o vetor normal ao hiperplano e $\cdot$ denota o produto interno.

Nesta etapa do trabalho surgiu o problema de desequilíbrio das classes, ou seja, do possível número muito diferente de amostras em cada classe. Na literatura existem duas maneiras de abordar o problema; a primeira implica o uso de algoritmos de reamostragem; a segunda propõe adaptar o algoritmo de aprendizado. Neste trabalho foram usadas ambas abordagens. Inicialmente foi usado o algoritmo SMOTE [Chawla *et al.*, 2002]. O problema com a abordagem inicial foi o tempo de

cômputo dos vetores de características porque devem ser calculados $\frac{n(n-1)}{2}$ vetores de características um para cada aresta do grafo $G(V, E)$ com n vértices como é mostrado na Figura 3.11 .

Figura 3.11: Representação gráfica do procedimento de criação dos vetores de características.

Na segunda etapa, o algoritmo de aprendizado foi adaptado. Neste trabalho, estamos interessados na predição de arestas que mudam de estado no tempo. Portanto, foram desconsideradas as arestas que se mantiveram constantes durante o período de tempo analisado. Com isso, foi eliminada a maioria das arestas pertencentes à classe negativa, porque nas redes sociais, a maioria das arestas não aparece, e nosso conjunto de dados ficou balanceado. A Figura 3.12 mostra graficamente o procedimento de adaptação do algoritmo, desconsiderando as arestas em preto.

Figura 3.12: Representação gráfica do procedimento de adaptação do algoritmo de aprendizado.

3.3.2 Descrição dos dados

Com a finalidade de avaliar o método de predição de ligações baseado em aprendizado supervisionado, foram realizados vários testes sobre os dados extraídos utilizando scripLattes [Mena-Chalco e Cesar-Jr, 2009] da Plataforma Lattes. Neste experimento, foram consideradas oito redes

Área de conhecimento	Pesquisadores
Astronomia	775
Ciência da computação	7677
Física	10572
Geociências	5790
Matemática	5667
Oceanografia	1415
Probabilidade e Estatística	2771
Química	12888

Tabela 3.4: *Quantidade de pesquisadores para as áreas de conhecimento consideradas.*

de colaboração científica reais formadas por doutores de todas as áreas da Grande Área: Ciências Exatas e da Terra. As informações dos pesquisadores foram extraídas da plataforma Lattes até, o 28 de outubro de 2013. Ao todo foram estudados 41 739 Currículos Lattes únicos; a Tabela 3.4 mostra as áreas consideradas e o número de pesquisadores avaliados para cada área. Adicionalmente, também foi analisado um grupo de 889 docentes vinculados a 45 programas de pós-graduação em Ciências da Computação no Brasil durante o período 2000 - 2012.

A seguir, são apresentadas graficamente estas redes de colaboração científica; também são apresentadas as componentes gigantes (maior componente conexa do grafo). Algumas medidas de centralidade usam apenas as componente gigantes perdendo com isto informações importantes. Em nosso caso, são considerados todos os vértices e arestas.

O vetor de características formado por atributos de vértices e arestas foi submetido a um processo de seleção de características utilizando SVM e F-score como descrito em [Chen e Lin, 2006].

As Figuras 3.21 e 3.22 mostram os *rankings* individuais de relevância de todas as características usando resultados dos testes com SVM. A primeira mostra o *ranking* usando redes ponderadas, enquanto a segunda mostra o *ranking* para redes simples não ponderadas. É possível observar na Figura 3.21, que as 5 primeiras características conseguem individualmente melhores resultados. Por outro lado, na Figura 3.22 as 4 primeiras características apresentam melhor desempenho.

A Figura 3.23 mostra o resultado depois de aplicar o algoritmo de seleção de características. Neste gráfico, é mostrado o desempenho do preditor com diversos números de características entre 1 e 11 para redes não ponderadas, e 1 e 12 para redes ponderadas. É possível observar que usando redes ponderadas, o melhor resultado foi obtido considerando-se 4 características, a saber: são caminho mais curto, número de vizinhos em comum, ligação preferencial e o índice Adamic/Adar. Para redes não ponderadas, por outro lado, as características escolhidas foram: o número de vizinhos em comum, ligação preferencial, o índice Adamic/Adar e o índice de Jaccard.

(a)

(b)

Figura 3.13: Rede de colaboração da astronomia para o período 2010-2012, contém 774 vértices e 1058 arestas; o componente gigante contém 374 (48,32%) vértices e 1033 (97,67%) arestas.

(a)

(b)

Figura 3.14: Rede de colaboração da ciência da computação no Brasil para o período 2010-2012, contém 7677 vértices e 7929 arestas; o componente gigante contém 3256 (42,41%) vértices e 7622 (96,13%) arestas.

(a)

(b)

Figura 3.15: Rede de colaboração da física no Brasil para o período 2010-2012, contém 10572 vértices e 17357 arestas; o componente gigante contém 5377 (50,86%) vértices e 16870 (97,19%) arestas.

(a)

(b)

Figura 3.16: Rede de colaboração da geociências no Brasil para o período 2010-2012, contém 5790 vértices e 7103 arestas; o componente gigante contém 2803 (48,41%) vértices e 16870 6863 (96,62%) arestas.

(a)

(b)

Figura 3.17: Rede de colaboração da matemática no Brasil para o período 2010-2012, contém 5667 vértices e 2386 arestas; o componente gigante contém 1044 (18,42%) vértices e 16870 1538 (64,46%) arestas.

(a)

(b)

Figura 3.18: Rede de colaboração da oceanografia no Brasil para o período 2010-2012, contém 1415 vértices e 1583 arestas; o componente gigante contém 710 (50,18%) vértices e 16870 1557 (98,36%) arestas.

(a)

(b)

Figura 3.19: Rede de colaboração da probabilidade no Brasil para o período 2010-2012, contém 2771 vértices e 1286 arestas; o componente gigante contém 542 (19,56%) vértices e 16870 984 (76,52%) arestas.

(a)

(b)

Figura 3.20: Rede de colaboração da química no Brasil para o período 2010-2012, contém 12888 vértices e 27612 arestas; o componente gigante contém 7793 (60,47%) vértices e 27372 (99,13%) arestas.

Figura 3.21: Desempenho individual das características consideradas no vetor do preditor, usando redes simples ponderadas

3.3.3 Resultados

Com as características escolhidas, a média das medidas de desempenho do preditor são apresentadas na Tabela 3.5.

Acurácia	Precisão	Sensibilidade	Medida-F
0,8684	0,7927	0,9143	0,8491

Tabela 3.5: Média dos valores obtidos da Acurácia, Precisão, Sensibilidade e Medida-F do preditor de ligação.

Em que um preditor p , que faz uma predição positiva (T) ou negativa (F), se a previsão positiva é correta, diz ser um verdadeiro positivo (TP), caso contrário, é um falso-positivo (FP). Da mesma forma, uma previsão negativa correta é um verdadeiro-negativo (TN), caso contrário, é um falso-negativo (FN). Como as classes não são completamente balanceadas, isso provoca um comportamento anormal da medida de precisão. Consideramos três medidas alternativas para avaliar o desempenho de previsão:

$$Acurcia = \frac{TP + TN}{TP + FP}$$

$$Precisão = \frac{TP}{TP + FP}$$

$$Sensibilidade = \frac{TP}{TP + FN}$$

$$medida - F = \frac{2 \times Precisão \times Sensibilidade}{Precisão + Sensibilidade}$$

Figura 3.22: Desempenho individual das características consideradas no vetor do preditor, usando redes ponderadas

3.4 Influência na rede de colaboração

O método proposto foi dividido em três etapas, as quais são apresentadas na Figura 3.24. A primeira etapa é identificar as redes de colaboração científica. Dadas as redes, o algoritmo de predição de ligação supervisionado é executado n vezes (n é o número de vértices da rede). A cada iteração remove-se um vértice diferente do conjunto de teste. Finalmente, todos os resultados das predições são ordenados para examinar a influência de colaboração de cada pesquisador com base em como sua presença ou ausência na rede afeta o resultado de predição de ligação.

Neste trabalho, são consideradas redes de colaboração não dirigidas tanto ponderadas como não ponderadas; foram construídas as redes conforme foi descrito na 3.3.

A predição de ligação pode ser modelada como um problema de classificação supervisionado em que cada ponto de dados corresponde a uma possível aresta (uma ligação entre um par de pesquisadores), como foi descrito na Seção 3.3.

A segunda fase do processo inicia-se com um pré-processamento das redes de colaboração para gerir a sua natureza dinâmica, colapsando um subconjunto de redes (correspondente a uma janela temporal) em um único grafo “resumo”. O método de união consiste em combinar as arestas atribuindo um peso que é determinado pela contribuição dada à aresta temporal nas redes “resumo”. Pesos maiores são atribuídos a trabalhos colaborativos mais recentes e pesos menores para colaborações mais antigas. Uma abordagem semelhante foi adotada por Sharan e Neville [2008] e Acar *et al.* [2009], nos quais as ligações são ponderadas para trás no tempo das redes anuais de X_1 para X_T , de acordo com a Equação (3.11).

Figura 3.23: Desempenho em grupo das características consideradas no vetor do preditor, (a) usando redes ponderadas, (b) usando redes simples

$$G'_x(i, j) = \sum_{t=1}^T (1 - \theta)(T - t)G_{xt}(i, j). \quad (3.11)$$

G' é a rede colapsada e $G_t(i, j)$ é uma possível aresta entre os vértices i e j no tempo t . θ é uma constante definida pelo usuário para administrar a duração da ligação ao longo do tempo e é definida após algumas experiências realizadas no conjunto de treinamento. Posteriormente, criamos os vetores de características para o algoritmo supervisionado usando as redes previamente colapsadas. Os recursos são calculados para cada par de vértices. Como trabalhamos com redes não dirigidas, um mesmo vetor de características descreve os pares de vértices (i, j) e (j, i) . Uma representação gráfica deste processo é apresentada na Figura 3.25.

Considerações levadas em conta nesta etapa foram semelhantes àsquelas descritas na Seção 3.3.

Na fase final, o preditor de ligações criado é testado com uma rede resumo ponderada criada para uma janela de tempo futura em relação à janela de treinamento. Este primeiro teste é chamado de “teste de rede completa”, e o resultado é tomado como referência para o cálculo da influência no resultado da predição de ligação. Em seguida, foi feita uma série de testes, um para cada pesquisador que pode ser removido da rede. Estes testes são chamados de “testes de sub-rede”. Cada teste consiste em remover da rede um pesquisador diferente (vértice) e suas respectivas ligações, a fim de obter uma evidência de como a ausência dessa pessoa afeta a previsão da ligação na rede. Em seguida, todos os resultados dos “testes de sub-rede” são comparados com os do “teste de rede completa”, a fim de avaliar em que medida a ausência de um pesquisador em particular afeta o desempenho de precisão da ligação, e se esta influência é positiva ou negativa. Os pesquisadores são classificados pela influência calculada, com o objetivo de identificar aqueles que têm a maior influência na rede como um todo, com base em seus resultados de previsão de ligação.

Seja R_f o conjunto de ligações de preditas pelo teste de rede completa e R_i $i \in [1..n]$ o conjunto

Figura 3.24: Representação esquemática do método proposto para detectar a influência de colaboração de pesquisadores em rede de colaboração científica.

de ligações preditas pelo teste arbitrário de sub-rede t_i .

Para uma rede não ponderada, a influência colaborativa sem sinal é medida como segue:

$$UCI(t_i) = |(R_f \Delta R_i)| \times (F\text{-measure}_{t_i}) \quad (3.12)$$

no qual $(R_f \Delta R_i)$ denota o conjunto de arestas que geram uma predição diferente no teste de rede completa e no teste de sub-rede t_i , $|X|$ denota a cardinalidade de um conjunto X e $F\text{-measure}_{t_i}$ denota a medida-F do preditor relativo ao teste t_i . Com a finalidade de avaliar se a ausência de um vértice produz um impacto positivo ou negativo no desempenho da predição de ligações a Equação 3.12 foi modificada para obter a medida de influência colaborativa com sinal a qual é calculada da seguinte forma:

$$SCI(t_i) = (|E| - |L|) \times (F\text{-measure}_{t_i}) \quad (3.13)$$

onde E representa o conjunto de arestas adquiridas e L é o conjunto de arestas perdidas do conjunto $(R_f \Delta R_i)$.

Para redes ponderadas, as duas equações foram modificadas para trabalhar com os pesos das arestas.

3.4.1 Descrição dos dados

Como ocorrido nos caso da predição de ligações, as redes de coautoria bibliográfica do período de 2000 a 2012, também foram extraídas da Plataforma Lattes do Brasil usando a ferramenta *scriptLattes* [Mena-Chalco e Cesar-Jr, 2009]. Foram usadas as mesmas redes de colaboração da Tabela

Figura 3.25: Representação gráfica do processo de criação do classificador. Primeiro, uma rede ponderada resumo é criada pela união de várias redes dentro de uma janela temporal. Subsequentemente, um conjunto de características é extraído a partir da rede de resumo e usados para treinar o classificador.

3.4 para fazer os testes; adicionalmente, uma rede formada por 889 pesquisadores vinculados aos programas pós-graduação em Ciência da Computação no Brasil.

3.4.2 Resultados

Na construção das redes de coautoria têm sido considerados os seguintes tipos de produção bibliográfica: (i) artigos em revistas científicas, (ii) Livro publicado/organizado, (iii) Capítulo de livro publicado, (iv) Artigos em jornais/revistas, (v) trabalhos completos publicados em anais de congressos, (vi) resumos expandidos publicados em anais de congressos, e (vii) resumos publicados em anais de congressos.

A Figura 3.26 ilustra o tipo de rede de colaboração analisada neste trabalho. Como observado anteriormente, as redes possuem um número fixo de vértices, enquanto que as arestas entre os vértices variam dinamicamente ao longo do tempo. Trabalhamos com redes reais em evolução descrevendo colaborações científicas entre pesquisadores afiliados a 45 programas de pós-graduação em Ciência da Computação no Brasil. Um total de 889 pesquisadores são considerados, por um

Figura 3.26: *Dinâmica das arestas na rede de colaboração dos programas de pós-graduação em Ciência da Computação no Brasil ao longo dos últimos 12 anos, com um número fixo de vértices. Os pesquisadores são identificados por um número de código único.*

período entre 2000 e 2012. Como pode ser observado na Figura 3.31, a rede possui várias componentes conexas e também vértices isolados. Isto pode ocorrer por várias razões, mas do ponto de vista da colaboração, os vértices isolados não devem exercer nenhuma influência na rede.

Dividimos os dados de acordo com o ano da atividade de colaboração, considerando na análise partições formadas por anos consecutivos e não sobrepostos.

Em primeiro lugar, foi gerada a rede de colaboração para cada ano, independentemente das partições. Em seguida, o método descrito previamente foi usado para obter a rede “resumo” de cada partição. Com base em experiências anteriores, adotamos $\theta = 0,2$ na Equação 3.11; para colapsar

Figura 3.27: Rede de colaboração de todos os pesquisadores afiliados a 45 programas de pós-graduação em Ciência Computação no Brasil, para o período de 2009 a 2012. Cada pesquisador é representado por um vértice, e cada trabalho colaborativo é representado por uma aresta da rede.

as redes foi atribuída maior relevância para as colaborações mais recentes. Assim, a partição correspondente aos primeiros anos foi tomada como conjunto de treinamento para o preditor, e a segunda partição foi usada para o teste.

Neste estudo, estamos especialmente interessados na dinâmica de criação das ligações nas redes de colaboração, ou seja, como as ligações estão sendo criadas e removidas. Assim, a classe alvo do vetor de treinamento é determinada através da detecção de mudanças no estado da ligação existente (para não-existente, ou vice-versa) a partir do treinamento para os anos de testes.

Para construir o preditor, precisamos calcular um vetor de características para cada possível aresta (pares de vértices) na rede, conforme descrito previamente. Neste ponto, observa-se que a percentagem de exemplos positivos é inferior a 10% do conjunto de possíveis arestas, então adotamos uma das melhores estratégias de sub amostragem *oversampling*, o SMOTE [Chawla *et al.*, 2002]. Assim, consideramos uma percentagem de 400% de aumento da classe com menor número de exemplos. Como este *oversampling* aumenta consideravelmente o tamanho do conjunto de treinamento, realizamos uma subamostragem de propagação. Estes passos de pré-processamento

reduzem o problema de desequilíbrio e também o tamanho do conjunto de treino.

Para o exemplo apresentado a seguir, uma amostra de pesquisadores vinculados a um dos programas de pós-graduação em Ciência da Computação foi selecionada, e os seus currículos Lattes foram processados de acordo com o procedimento descrito acima.

O período de estudo foi dividido em três intervalos de tempo, sendo que o primeiro corresponde aos anos 2000 e 2006, o segundo, aos anos 2007 a 2009, e o terceiro, aos anos de 2010 a 2012. As redes de colaboração correspondentes ao primeiro período foram unidas em uma rede. Para estas redes foram obtidos os vetores de características e utilizados como conjunto de treinamento para o preditor.

A rede de colaboração do segundo intervalo de tempo foi usada para determinar a classe dos vetores de características no conjunto de treinamento, e foi usada para calcular as características dos vetores do conjunto de teste para prever as ligações para a rede resumida do terceiro período (de 2010 a 2012). Para esta rede, calculamos o *ranking* da influência de colaboração de cada vértice nas redes tanto ponderada como não ponderada. No total, 40 testes foram realizados sendo um teste de rede completa e 39 testes de sub-rede, um para cada vértice removível. A Figura 3.28 mostra os resultados.

O gráfico da Figura 3.28a foi criado com base no *ranking* de resultados para influência colaborativa sem sinal estimada a partir das redes não ponderadas (ICSSRN). Os vértices que, ao serem removidos da rede, causaram maior variação nos resultados da predição de ligação estão pintados em preto, e aqueles que causaram menor variação estão pintados de branco. De forma semelhante, a Figura 3.28c mostra o *ranking* da influência na colaboração sem sinal sobre redes ponderadas (ICSSRP). Percebe-se que a influência de colaboração tem alguma correlação positiva com a *Degree centrality*. A correlação Kendall calculada foi 0,77 e 0,58. Também calculamos a correlação com outras medidas de centralidade conhecidas, conforme ilustrado na Figura 3.29. A escala de cores enfatiza as correlações positivas e negativas em tons de laranja, vermelho e azul-violeta, respectivamente.

A ICSSRN (*Influência Colaborativa com Sinal em Redes Não Ponderadas*) corresponde à correlação do *ranking* proposto com todas as outras métricas comparáveis, calculado a partir dos mesmos recursos.

As redes apresentadas na 3.28b e 3.28d, por outro lado, foram criadas com base no *ranking* da influência colaborativa com sinal estimada a partir das redes não ponderadas e ponderadas respectivamente. Neste caso, observa-se que a remoção dos vértices centrais provoca maior variação negativa nos resultados da predição da ligação. Esses vértices são pintados em vermelho, enquanto que aqueles que geraram um efeito positivo no desempenho da predição são apresentados em azul. Novamente, calculamos a correlação com outras medidas de centralidade conhecidas, como apresentado na Figura 3.29. Neste caso, um valor de correlação de Kendall -0,76 é obtido com a centralidade de grau (*Degree centrality*) e nossa medida sobre redes ponderadas e um valor de -0,67 é obtido com a centralidade por intermediação (*Betweenness centrality*). Por outro lado, uti-

Figura 3.28: (a) ICSSRN e (c) ICSSRP mostram o ranking da influência na colaboração sem sinal, usado em rede não ponderada e ponderada respectivamente, os vértices pretos e maiores representam aos indivíduos mais influentes, enquanto o branco e menores são os menos influentes. (b) ICSSRN e (d) ICSSRP mostram o ranking da influência colaborativa usando rede não ponderada e ponderada respectivamente, os vértices em vermelho são os mais influentes, vértice cuja remoção afeta negativamente o desempenho da predição de ligações. Os vértices azuis são os que, ao serem removidos, causam uma melhora no desempenho da predição. Os valores intermediários, com baixa influência na colaboração, são representados em cor branca.

lizando redes não ponderadas, a correlação obtida foi de -0.74 com a centralidade de grau (*Degree centrality*) e -0.75 com a centralidade por intermediação (*Betweenness centrality*).

No segundo exemplo ilustrativo, examinamos 889 pesquisadores vinculados a 45 programas de pós-graduação em Ciência da Computação no Brasil. A Figura 3.30 apresenta a evolução das

Figura 3.29: Correlação entre as métricas centralidade conhecidas e a medida proposta (ICRNSS, ICRNCS, ICRPSS, e ICRPCS)

colaborações entre esses pesquisadores no tempo. A rede na Figura 3.30a corresponde ao período entre 2007-2009, e foi usado como conjunto de teste para prever as ligações da rede apresentada na 3.30b, correspondente ao período entre 2010-2012. Neste caso uma rede corresponde ao período entre 2004 e 2006, foi usada para treinar o preditor.

A Figura 3.31 mostra uma representação visual do *ranking* da influência dos vértices na rede de colaboração usando um modelo de redes ponderadas. Na 3.31a, é apresentada a versão sem sinal da influência colaborativa (ICRPSS), destacando os vértices mais influentes na cor preta, enquanto os menos influentes são apresentados em cor branca.

O *ranking* da influência colaborativa com sinal (ICRPCS) é apresentado na Figura 3.31b, na qual os vértices vermelhos representam os indivíduos mais influentes, cuja ausência na rede afeta negativamente o desempenho da predição de ligações. Por outro lado, a remoção dos vértices azuis melhora o desempenho da predição de ligações. Os vértices neutrais são apresentados na cor branca.

Essa figura permite identificar os pesquisadores responsáveis pela evolução da rede do estado mostrado na Figura 3.30a para o estado mostrado na Figura 3.30b, em termos de colaborações.

Assim como no primeiro exemplo apresentado, foi observada a existência de uma correlação de pelo menos -0,58 entre a medida proposta e as medidas de centralidades de grau, intermediação e proximidade.

Figura 3.30: Rede de colaboração de todos os pesquisadores vinculados a 45 programas de pós-graduação de Ciências da Computação no Brasil, para o período(a) 2007-2009, (b) 2010-2012. Os vértices representam pesquisadores e as arestas indicam relações de coautoria.

Figura 3.31: Ranking da influência colaborativa de cada pesquisador em uma rede de colaboração ponderada; a) ICRPSS, os vértices maiores e pretos representam os pesquisadores mais influentes, enquanto os menores e brancos representam os pesquisadores menos influentes; (b) ICRPCS, a ausência dos vértices vermelhos afeta negativamente o desempenho da predição de ligações, enquanto a ausência dos vértices azuis afetam positivamente a predição.

Capítulo 4

Considerações Finais

Neste capítulo são apresentadas as considerações finais do trabalho. São sugeridos alguns trabalhos futuros e, em seguida, são citadas as publicações resultantes desta pesquisa.

A fim de testar o modelo proposto, consideramos 9 conjuntos de pesquisadores cadastrados na plataforma Lattes. Ao todo foram considerados 47555 pesquisadores diferentes. As redes de colaboração foram inicialmente criadas utilizando uma adaptação do script Lattes. Essas redes foram posteriormente enriquecidas com os atributos descritivos dos pesquisadores tais como número de publicações individuais em diferentes categorias e número de publicações internacionais. Como atributos de arestas foram considerados os trabalhos colaborativos distribuídos nas seguintes categorias: (i) artigos em revistas científicas, (ii) Livro publicado/organizado, (iii) Capítulo de livro publicado, (iv) Artigos em jornais/revistas, (v) trabalhos completos publicados em anais de congressos, (vi) resumos expandidos publicados em anais de congressos, e (vii) resumos publicados em anais de congressos.

O modelo proposto, usando grafos relacionais com atributos, permitiu caracterizar de maneira flexível as redes de colaboração científica. Permitiu também a automação na: (i) criação, (ii) enriquecimento, (iii) implementação de medidas estruturais locais e globais, (iv) facilidade de exportação de dados e (v) análise de grupos de pesquisadores. Esse modelo admite em um único arcabouço informação descritivas dos pesquisadores e estruturais dos elementos das redes de colaboração. Dependendo da aplicação essas informações são importantes em grau diferente.

Os resultados apresentados neste trabalho permitem obter uma radiografia dos grupos de pesquisa estudados, em termos de produção bibliográfica, interação/inserção acadêmica, e grau de internacionalização. Essa informação pode ser considerada para avaliar o desempenho/desenvolvimento dos grupos, detectar as sub-comunidades de interação acadêmica, identificar possíveis carências, e selecionar potenciais parcerias com pesquisadores externos.

Baseado em nossos resultados, o método de cálculo da taxa de internacionalização é considerado robusto na identificação de países nos endereços dos autores de produções bibliográficas. Conseguimos determinar que as características vinculadas com as arestas são as mais relevantes

para a predição de ligações. Neste sentido para redes ponderadas as características escolhidas foram distância entre dois vértices, número de vizinhos em comum, coeficiente de Jaccard e Índice Adamic/Adar. Por outro lado para redes não ponderadas as características escolhidas foram número de vizinhos em comum, coeficiente de Jaccard, Índice Adamic/Adar e ligação preferencial.

As principais contribuições deste trabalho, além do próprio arcabouço para modelar e analisar redes de colaboração científica com ARGs, incluem as medidas de influência colaborativa dos pesquisadores em uma rede de colaboração científica, quantificando a influência com base em como a presença ou ausência dos pesquisadores afeta o desempenho da predição de ligação na rede. Estas medidas foram definidas em termos de medidas estruturais principalmente locais, que têm sido empregadas para treinar um classificador SVM utilizado como preditor. Os resultados da predição de ligações considerando a rede completa e a rede sem o pesquisador alvo são comparadas considerando a medida-F e a variação resultante sobre as ligações e os pesos das ligações previstas, a fim de determinar a influência do pesquisador na rede de colaboração.

As medidas propostas da influência colaborativa com sinal apresentaram correlação inversa significativa com as medidas de centralidade mais conhecidas, o que confirma a efetividade das medidas propostas. Além disso, estas medidas podem ser calculadas independentemente para cada vértice e não exigem o cálculo de medidas globais

4.0.3 Trabalhos futuros

Diversas linhas podem ser exploradas a partir do trabalho descrito nesta dissertação. O modelo proposto deve ser avaliado de maneira mais ampla usando redes de pesquisadores maiores assim como redes com maior quantidade de atributos. Também deve ser criado um mecanismo de comparação da medida de influência proposta neste trabalho com medidas semelhantes mais recentes, que garanta a equidade da comparação. Além disso existem diversas aplicações que podem ser exploradas usando o modelo proposto.

- É possível usar os resultados da influência colaborativa em diferentes níveis hierárquicos com a finalidade de determinar a existência de algum tipo de correlação entre os resultados. Este estudo permitiria determinar se pesquisadores influentes em redes menores ao ser avaliados em redes maiores também são influentes.
- De maneira semelhante à proposta para cálculo da influência colaborativa individual é possível agrupar pesquisadores e determinar a influência colaborativa de um grupo de pesquisadores em grandes redes. Assim acreditamos que o modelo proposto possa ser útil para investigar a influência colaborativa de subgrupos.
- Estudos mais detalhados baseados na ausência de pesquisadores numa rede também podem ser feitos com a finalidade de determinar o impacto de aposentadorias ou mudanças de emprego de pesquisadores.

A precisão do método de predição de ligações pode ser melhorada através do uso de características mais específicas ou adaptação do algoritmo de aprendizado. Neste estudo foram usadas principalmente medidas de locais nos vetores de características, mas poderiam ser usadas medidas globais também.

4.0.4 Publicacoes

Como produto desta pesquisa, foram publicados três artigos relatando os estudos na área

- O primeiro deles [Perez-Cervantes et al., 2012a] foi apresentado no "2nd Workshop on Analyzing and Improving Collaborative eScience with Social Networks" em Chicago, IL, EUA. Esse artigo descreve o método automatizado para o cálculo da taxa de internacionalização descrito na seção Seção 3.2 do Capítulo 3 .
- O segundo artigo [Perez-Cervantes et al., 2012b] foi apresentado no "VI Simpósio de Iniciação Científica e Pós-graduação IME-USP" o qual contém um estudo da interação acadêmica dos docentes do IME-USP entre os anos 2000 e 2011.
- O terceiro artigo [Perez-Cervantes et al., 2013] contendo um estudo da influência individual na rede de colaboração dos pesquisadores de ciências da computação do Brasil descrito na Seção 3.4 do Capítulo 3, foi aceito na 2013 IEEE 9th International Conference on e-Science (e-Science 2013) em Beijing, China.

Bibliografia

- Abbasi(2013)** Alireza Abbasi. h-type hybrid centrality measures for weighted networks. *Scientometrics (In press)*. Citado na pág. 19
- Abbasi et al.(2010)** Alireza Abbasi, Jörn Altmann e Junseok Hwang. Evaluating scholars based on their academic collaboration activities: two indices, the rc-index and the cc-index, for quantifying collaboration activities of researchers and scientific communities. *Scientometrics*, 83(1): 1–13. Citado na pág. 19
- Acar et al.(2009)** Evrim Acar, Daniel M Dunlavy e Tamara G Kolda. Link prediction on evolving data using matrix and tensor factorizations. Em *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, páginas 262–269. IEEE. Citado na pág. 17, 18, 51
- Adamic e Adar(2003)** Lada A Adamic e Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230. Citado na pág. 5, 16, 18, 38
- Al Hasan e Zaki(2011)** Mohammad Al Hasan e Mohammed J Zaki. A survey of link prediction in social networks. Em *Social network data analytics*, páginas 243–275. Springer. Citado na pág. 1, 15
- Al Hasan et al.(2006)** Mohammad Al Hasan, Vineet Chaoji, Saeed Salem e Mohammed Zaki. Link prediction using supervised learning. Em *SDM'06: Workshop on Link Analysis, Counter-terrorism and Security*. Citado na pág. 15, 16
- Barabási e Albert(1999)** Albert-László Barabási e Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512. Citado na pág. 10, 16, 18
- Bilgic et al.(2007)** Mustafa Bilgic, Galileo Mark Namata e Lise Getoor. Combining collective classification and link prediction. Em *Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on*, páginas 381–386. IEEE. Citado na pág. 16
- Bonacich(1987)** Phillip Bonacich. Power and centrality: A family of measures. *American journal of sociology*, páginas 1170–1182. Citado na pág. 21
- Bondy e Murty(1976)** John Adrian Bondy e Uppaluri Siva Ramachandra Murty. *Graph theory with applications*, volume 6. Macmillan London. Citado na pág. 4
- Bornholdt e Schuster(2006)** Stefan Bornholdt e Heinz Georg Schuster. *Handbook of Graphs and Networks*. Citado na pág. 9

- Burges(1998)** Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167. Citado na pág. 39
- Chawla et al.(2002)** Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall e W Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357. Citado na pág. 39, 56
- Chen e Lin(2006)** Yi-Wei Chen e Chih-Jen Lin. Combining svms with various feature selection strategies. Em *Feature extraction*, páginas 315–324. Citado na pág. 41
- Clauset et al.(2008)** Aaron Clauset, Cristopher Moore e Mark EJ Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101. Citado na pág. 17
- Davis et al.(2011)** Darcy Davis, Ryan Lichtenwalter e Nitesh V. Chawla. Multi-relational link prediction in heterogeneous information networks. Em *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '11, páginas 281–288, Washington, DC, USA. IEEE Computer Society. ISBN 978-0-7695-4375-8. Citado na pág. 17
- de Sá e Prudêncio(2011)** Hially Rodrigues de Sá e Ricardo Bastos Cavalcante Prudêncio. Supervised link prediction in weighted networks. Em *Neural Networks (IJCNN), The 2011 International Joint Conference on*, páginas 2281–2288. IEEE. Citado na pág. 17
- Dolby e Rahman(2008)** N. Dolby e A. Rahman. Research in international education. *Review of Educational Research*, 78(3):676–726. Citado na pág. 12
- Doppa et al.(2009)** Janardhan Rao Doppa, Jun Yu, Prasad Tadepalli e Lise Getoor. Chance-constrained programs for link prediction. Em *Proceedings of workshop on analyzing networks and learning with graphs at NIPS conference*. Citado na pág. 16
- Dunlavy et al.(2011)** Daniel M. Dunlavy, Tamara G. Kolda e Evrim Acar. Temporal link prediction using matrix and tensor factorizations. *TKDD*, 5(2):10. Citado na pág. 17
- Estrada e Hatano(2008)** Ernesto Estrada e Naomichi Hatano. Communicability in complex networks. *Physical Review E*, 77(3):036111. Citado na pág. 19
- Estrada e Rodriguez-Velazquez(2005)** Ernesto Estrada e Juan A Rodriguez-Velazquez. Subgraph centrality in complex networks. *Physical Review E*, 71(5):056103. Citado na pág. 19
- Freeman(1977)** Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, páginas 35–41. Citado na pág. 18, 19, 20
- Freeman(1979)** Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239. Citado na pág. 20
- Getoor et al.(2003)** Lisa Getoor, Nir Friedman, Daphne Koller e Benjamin Taskar. Learning probabilistic models of link structure. *The Journal of Machine Learning Research*, 3:679–707. Citado na pág. 17
- Getoor e Diehl(2005)** Lise Getoor e Christopher P Diehl. Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, 7(2):3–12. Citado na pág. 1

- Girvan e Newman(2002)** Michelle Girvan e Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826. Citado na pág. 10
- Glänzel et al.(2006)** W. Glänzel, J. Leta e B. Thijs. Science in brazil. part 1: A macro-level comparative study. *Scientometrics*, 67(1):67–86. Citado na pág. 14, 26
- Gong et al.(2011)** Neil Zhenqiang Gong, Ameet Talwalkar, Lester Mackey, Ling Huang, Eui Chul Richard Shin, Emil Stefanov, Dawn Song et al. Jointly predicting links and inferring attributes using a social-attribute network (san). Citado na pág. 1, 10, 17, 25
- Guimera et al.(2006)** Roger Guimera, Marta Sales-Pardo e Luís AN Amaral. Classes of complex networks defined by role-to-role connectivity profiles. *Nature physics*, 3(1):63–69. Citado na pág. 17
- Han(2009)** Jiawei Han. Mining heterogeneous information networks by exploring the power of links. Em *Proceedings of the 12th International Conference on Discovery Science*, DS '09, páginas 13–30, Berlin, Heidelberg. Springer-Verlag. ISBN 978-3-642-04746-6. Citado na pág. 2
- Hassan e Haddawy(2012)** S. Hassan e P. Haddawy. Measuring international knowledge flows and scholarly impact of scientific research. *Scientometrics*, páginas 1–17. Citado na pág. 13
- Heckerman et al.(2004)** David Heckerman, Christopher Meek e Daphne Koller. Probabilistic models for relational data. Relatório técnico, Technical Report MSR-TR-2004-30, Microsoft Research. Citado na pág. 17
- Hirsch(2005)** Jorge E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102(46): 16569. Citado na pág. 19
- Hu et al.(2012)** Ke Hu, Ju Xiang, Wanchun Yang, Xiaoke Xu e Yi Tang. Link prediction in complex networks by multi degree preferential-attachment indices. *arXiv preprint arXiv:1211.1790*. Citado na pág. 15, 16, 18
- Kalviainen e Oja(1990)** Heikki Kalviainen e Erkki Oja. Comparisons of attributed graph matching algorithms for computer vision. Em *In Proc. of STEP-90, Finnish Artificial Intelligence Symposium*. Citeseer. Citado na pág. 12
- Kashima e Abe(2006)** Hisashi Kashima e Naoki Abe. A parameterized probabilistic model of network evolution for supervised link prediction. Em *Data Mining, 2006. ICDM'06. Sixth International Conference on*, páginas 340–349. IEEE. Citado na pág. 17
- Katz e Hicks(1997)** J. Katz e D. Hicks. How much is a collaboration worth? a calibrated bibliometric model. *Scientometrics*, 40:541–554. Citado na pág. 13
- Katz(1953)** Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43. Citado na pág. 16
- Kim(2006)** K.-W. Kim. Measuring international research collaboration of peripheral countries: Taking the context into consideration. *Scientometrics*, 66:231–240. Citado na pág. 13, 14

- Knight(2004)** J. Knight. Internationalization remodeled: Definition, approaches, and rationales. *Journal of Studies in International Education*, 8(1):5–31. Citado na pág. 12, 13
- Leite et al.(2011)** P. Leite, R. Mugnaini e J. Leta. A new indicator for international visibility: exploring brazilian scientific community. *Scientometrics*, 88:311–319. Citado na pág. 13, 26
- Leta e Chaimovich(2002)** J. Leta e W. Chaimovich. Recognition and international collaboration: the brazilian case. *Scientometrics*, 53(3):325–335. Citado na pág. 14, 32
- Leta et al.(2006)** Jacqueline Leta, Wolfgang Glänzel e Bart Thijs. Science in brazil. part 2: Sectoral and institutional research profiles. *Scientometrics*, 67(1):87–105. Citado na pág. 14, 26, 30
- Liben-Nowell e Kleinberg(2003)** David Liben-Nowell e Jon Kleinberg. The link-prediction problem for social networks. Em *Conference on Information and Knowledge Management (CIKM'03)*, páginas 556–559. Citado na pág. 18
- Lichtenwalter et al.(2010)** Ryan N Lichtenwalter, Jake T Lussier e Nitesh V Chawla. New perspectives and methods in link prediction. Em *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, páginas 243–252. ACM. Citado na pág. 16
- Lü e Zhou(2009)** Linyuan Lü e Tao Zhou. Role of weak ties in link prediction of complex networks. Em *Proceedings of the 1st ACM international workshop on Complex networks meet information & knowledge management*, páginas 55–58. ACM. Citado na pág. 17, 18
- Lü e Zhou(2010)** Linyuan Lü e Tao Zhou. Link prediction in weighted networks: The role of weak ties. *EPL (Europhysics Letters)*, 89(1):18001. Citado na pág. 10
- Lü e Zhou(2011)** Linyuan Lü e Tao Zhou. Link prediction in complex networks: A survey. *Physica A Statistical Mechanics and its Applications*, 390:1150–1170. Citado na pág. 1, 15
- Maucher et al.(2011)** Markus Maucher, Barbara Kracher, Michael Kühl e Hans A Kestler. Inferring boolean network structure via correlation. *Bioinformatics*, 27(11):1529–1536. Citado na pág. 18
- Mena-Chalco e Cesar-Jr.(2011)** J. P. Mena-Chalco e R. M. Cesar-Jr. Towards automatic discovery of co-authorship networks in the brazilian academic areas. Em *IEEE Seventh International Conference on e-Science Workshops*, páginas 53–60. Citado na pág. 27
- Mena-Chalco e Cesar-Jr(2009)** Jesús P. Mena-Chalco e R. M. Cesar-Jr. scriptlattes: an open-source knowledge extraction system from the lattes platform. *J. Braz. Comp. Soc.*, 15(4):31–39. Citado na pág. iv, 23, 27, 40, 53
- Menon e Elkan(2011)** Aditya Menon e Charles Elkan. Link prediction via matrix factorization. *Machine Learning and Knowledge Discovery in Databases*, páginas 437–452. Citado na pág. 17
- Neto(2010)** I. R. Neto. Prospectiva da pós-graduação no brasil (2008 - 2022). Citado na pág. 1
- Newman(2001)** Mark EJ Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132. Citado na pág. 19

- Newman(2003)** Mark EJ Newman. Mixing patterns in networks. *Physical Review E*, 67(2): 026126. Citado na pág. 6, 10
- Newman e Park(2003)** ME Newman e Juyong Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3):36122. Citado na pág. 6, 10, 18
- Opsahl et al.(2010)** Tore Opsahl, Filip Agneessens e John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251. Citado na pág. 18, 19
- Perez-Cervantes et al.(2012)** E. Perez-Cervantes, J. P. Mena-Chalco e R. M. Cesar-Jr. Towards a quantitative academic internationalization assessment of brazilian research groups. Em *2nd workshop on Analyzing and Improving Collaborative eScience with Social Networks - IEEE eScience - 2012*, páginas 1–8. Citado na pág. 13
- Popescul e Ungar(2003)** Alexandrin Popescul e Lyle H Ungar. Statistical relational learning for link prediction. Em *IJCAI workshop on learning statistical models from relational data*, volume 2003. Citado na pág. 17
- Potgieter et al.(2009)** Anet Potgieter, Kurt A. April, Richard J. E. Cooke e Isaac O. Osunmakinde. Temporality in link prediction: Understanding social complexity. *E:CO*, 11. Citado na pág. 16, 18
- Raan(1998)** A. F. J. Van Raan. The influence of international collaboration on the impact of research results. *Scientometrics*, 42:423–428. Citado na pág. 13
- Sabidussi(1966)** Gert Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603. Citado na pág. 20
- Sanfeliu e King-Sun(1983)** A. Sanfeliu e F. King-Sun. A distance measure between attributed relational graphs for pattern recognition. *IEEE transactions on systems, man, and cybernetics*, 13(3):353–362. Citado na pág. 10
- Sharan e Neville(2008)** Umang Sharan e Jennifer Neville. Temporal-relational classifiers for prediction in evolving domains. Em *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, páginas 540–549. IEEE. Citado na pág. 51
- Shaw(1964)** Marvin E Shaw. Communication networks. *Advances in experimental social psychology*, 1:111–147. Citado na pág. 20
- Shin et al.(2012)** Donghyuk Shin, Si Si e Inderjit S. Dhillon. Multi-scale link prediction. Em *Proceedings of the 21st ACM international conference on Information and knowledge management, CIKM '12*, páginas 215–224, New York, NY, USA. ACM. ISBN 978-1-4503-1156-4. Citado na pág. 17
- Ströele et al.(2013)** Victor Ströele, Geraldo Zimbrão e Jano M. Souza. Group and link analysis of multi-relational scientific social networks. *Journal of Systems and Software*. ISSN 0164-1212. Citado na pág. 10, 15, 17
- Sun et al.(2011)** Yizhou Sun, Rick Barber, Manish Gupta, Charu C. Aggarwal e Jiawei Han. Co-author relationship prediction in heterogeneous bibliographic networks. Em *ASONAM*, páginas 121–128. Citado na pág. 10, 17

- Sun et al.(2012)** Yizhou Sun, Jiawei Han, Charu C. Aggarwal e Nitesh V. Chawla. When will it happen?: relationship prediction in heterogeneous information networks. Em *Proceedings of the fifth ACM international conference on Web search and data mining, WSDM '12*, páginas 663–672, New York, NY, USA. ACM. ISBN 978-1-4503-0747-5. Citado na pág. 17, 18
- Szell et al.(2010)** Michael Szell, Renaud Lambiotte e Stefan Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31):13636–13641. Citado na pág. 10
- Taskar et al.(2002)** Ben Taskar, Pieter Abbeel e Daphne Koller. Discriminative probabilistic models for relational data. Em *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, páginas 485–492. Morgan Kaufmann Publishers Inc. Citado na pág. 17
- Taskar et al.(2003)** Ben Taskar, Ming-Fai Wong, Pieter Abbeel e Daphne Koller. Link prediction in relational data. Em *Advances in neural information processing systems*, página None. Citado na pág. 17
- Tomita(1991)** Masaru Tomita. *Current issues in parsing technology*. Kluwer Academic Publishers, Boston. Citado na pág. 28
- Tsai e Fu(1979)** Wen-Hsiang Tsai e King-Sun Fu. Error-correcting isomorphisms of attributed relational graphs for pattern analysis. *Systems, Man and Cybernetics, IEEE Transactions on*, 9 (12):757–768. ISSN 0018-9472. doi: 10.1109/TSMC.1979.4310127. Citado na pág. 2, 11
- Vapnik(1999)** Vladimir Vapnik. *The nature of statistical learning theory*. springer. Citado na pág. 39
- Wang et al.(2007)** Chao Wang, Venu Satuluri e Srinivasan Parthasarathy. Local probabilistic models for link prediction. Em *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, páginas 322–331. IEEE. Citado na pág. 17
- Wang e Groth(2010)** Shenghui Wang e Paul Groth. Measuring the dynamic bi-directional influence between content and social networks. Em *The Semantic Web–ISWC 2010*, páginas 814–829. Springer. Citado na pág. 19
- Wasserman(1994)** Stanley Wasserman. *Social network analysis: Methods and applications*, volume 8. Cambridge university press. Citado na pág. 1
- Wasserman(1998)** Stanley Wasserman. *Social network analysis : methods and applications*. Cambridge [England] ; New York : Cambridge University; Press. Citado na pág. 2, 4, 19
- Watts e Strogatz(1998)** Duncan J Watts e Steven H Strogatz. Collective dynamics of 'small-world' networks. *nature*, 393(6684):440–442. Citado na pág. 9
- Wei et al.(2013)** Daijun Wei, Xinyang Deng, Xiaoge Zhang, Yong Deng e Sankaran Mahadevan. Identifying influential nodes in weighted networks based on evidence theory. *Physica A Statistical Mechanics and its Applications*, 392:2564–2575. Citado na pág. 19
- Yan e Gregory(2012)** Bowen Yan e Steve Gregory. Detecting community structure in networks using edge prediction methods. *Journal of Statistical Mechanics: Theory and Experiment*, 2012 (09):P09008. Citado na pág. 16

- Yan et al.(2013)** Xiangbin Yan, Li Zhai e Weiguo Fan. C-index: A weighted network node centrality measure for collaboration competence. *Journal of Informetrics*, 7(1):223–239. Citado na pág. 19
- Yang et al.(2011)** Shuang-Hong Yang, Bo Long, Alex Smola, Narayanan Sadagopan, Zhaohui Zheng e Hongyuan Zha. Like like alike: joint friendship and interest propagation in social networks. Em *Proceedings of the 20th international conference on World wide web*, páginas 537–546. ACM. Citado na pág. 1, 10, 26
- Yang et al.(2012)** Zeyao Yang, Damou Fu, Yutian Tang, Yongbo Zhang, Yunsheng Hao, Chen Gui, Xu Ji e Xin Yue. Link prediction based on weighted networks. Em Tianyuan Xiao, Lin Zhang e Minrui Fei, editors, *AsiaSim 2012*, Communications in Computer and Information Science, páginas 119–126. Springer Berlin Heidelberg. ISBN 978-3-642-34389-6. Citado na pág. 10, 15
- Yu et al.(2007)** Kai Yu, Wei Chu, Shipeng Yu, Volker Tresp e Zhao Xu. Stochastic relational models for discriminative link prediction. *Advances in neural information processing systems*, 19:1553. Citado na pág. 17
- Yu et al.(2012)** Xiao Yu, Quanquan Gu, Mianwei Zhou e Jiawei Han. Citation prediction in heterogeneous bibliographic networks. Em *SDM*, páginas 1119–1130. Citado na pág. 17
- Zhai et al.(2013)** Li Zhai, Xiangbin Yan e Bin Zhu. The h l-index: improvement of h-index based on quality of citing papers. *Scientometrics*, páginas 1–11. Citado na pág. 19
- Zheleva et al.(2010)** Elena Zheleva, Lise Getoor, Jennifer Golbeck e Ugur Kuter. Using friendship ties and family circles for link prediction. Em *Advances in Social Network Mining and Analysis*, páginas 97–113. Springer. Citado na pág. 10
- Zhenqiang G et al.(2012)** N. Zhenqiang G, W. Xu, L Huang, P. Mittal, E. Stefanov, V. Sekar e D. Song. Evolution of social-attribute networks: Measurements, modeling, and implications using google+. Citado na pág. 1
- Zhou et al.(2009a)** Tao Zhou, Linyuan Lü e Yi-Cheng Zhang. Predicting missing links via local information. *The European Physical Journal B*, 71(4):623–630. Citado na pág. 18
- Zhou et al.(2009b)** Yang Zhou, Hong Cheng e Jeffrey Xu Yu. Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment*, 2(1):718–729. Citado na pág. 1, 10, 17, 26
- Zhuge(2006)** H. Zhuge. Discovery of knowledge flow in science. *Communications of ACM*, 49(5):101–107. Citado na pág. 13